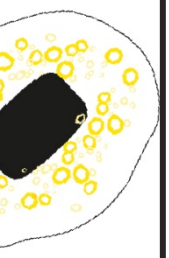

EREDIVISIE PLAYER VALUATION MODELS THROUGH THE APPLICATION OF MACHINE LEARNING METHODOLOGIES

UNIVERSITY OF TWENTE.

Master of Science - Business Administration - Thesis
Thom Treurniet 3082253

Examination committee
Dr. M. de Visser
Dr. M. Ehrenhard

University of Twente
Department of Behavioural, Management & Social Sciences
Master of Science - Business Administration
Specialisation - Digital Business and Analytics

Date: 10.09.2024
Wordcount: 14.175

Acknowledgements

This master's thesis fulfills the requirements for the Master of Science in Business Administration with a specialization in Digital Business and Analytics at the University of Twente. The development of Eredivisie player valuation models through the application of machine learning methodologies reflects my passion for football analytics and the innovative potential of machine learning in enhancing sports-related decision-making.

The journey leading to this research began with my growing interest in how data can be leveraged to improve the evaluation and management of football talent. Throughout my master's program I was fortunate to explore advanced machine learning techniques in courses such as Data Science. This naturally aligned with my fascination for football which inspired the focus of this study.

I am deeply grateful to Stefano Leone who played a pivotal role in this project by providing access to his SOFIFA EA24 dataset. His generosity in sharing this data was essential to the successful development of the models discussed in this thesis. I would also like to express my gratitude to Dr. M. de Visser and Dr. M. Ehrenhard for their supervision, support and valuable feedback throughout this research. Their critical thinking and encouragement played a pivotal role in shaping and finetuning my research. This led to a more thorough way of thinking and a more refined approach to achieving my research goals.

Additionally, I wish to acknowledge ChatGPT 4o for its significant contribution in aiding the data-preparation and model-building efforts in RStudio. This tool played a crucial role in accelerating my learning curve and allowed me to develop 10 player valuation models efficiently within this limited time frame.

This thesis offered a unique opportunity to blend theoretical research with practical applications in football analytics. The research experience has been both challenging and rewarding, equipping me with valuable insights that I will carry into my professional life. I hope this thesis can purpose as a foundation for further research in the field of football analytics.

Abstract

The valuation of football players has significant implications for professional football clubs because it influences their transfer decisions and long-term success. This thesis addresses the limitations of existing player valuation models by developing a more comprehensive and more accurate approach using machine learning methodologies. The research focuses on the Eredivisie and applies advanced machine learning techniques to create position-specific models that integrate a wide range of features. Features originate from categories such as player characteristics, performance, crowd-judgment, player potential and team features.

The study was conducted in two phases. The first phase involved identifying the influence of different features on player valuations through extensive feature analysis. The second phase focused on building and testing various machine learning models including Linear Regression, Ridge Regression, Lasso Regression, Principal Component Regression, Partial Least Squares Regression, Particle Swarm Optimization with Support Vector Regression, LightGBM, XGBoost, CatBoost and a Meta-model through ensemble stacking.

The findings reveal that models incorporating features from relevant subsets outperform those based on single subsets. Specifically, the Particle Swarm Optimization combined with Support Vector Regression model demonstrated superior performance for predicting the valuations of attackers and midfielders. The CatBoost model with Bayesian Optimization excelled in value prediction for defenders and the XGBoost model with Bayesian Optimization was most effective in value prediction for goalkeepers.

The study concludes that the best-performing position-based player valuation models built with the newest machine learning methods significantly outperform existing models, offering more precise and reliable predictions. The study thereby contributes to the academic field by advancing the integration of the latest machine learning techniques in football player valuation. Previous research often focusses on limited feature subsets or generic modeling approaches. This study demonstrates that a comprehensive position-based approach using the latest machine learning algorithms significantly enhances predictive accuracy. By incorporating features from relevant categories and leveraging the latest machine learning methods such as PSO-SVR, CatBoost and XGBoost, the study addresses the research gap related to feature integration and model specificity. These findings extend the theoretical understanding of player valuation models by providing evidence of the advantages of ensemble and optimization-based approaches, setting a new standard for future research in football analytics.

These prediction models also provide football clubs with more precise and actionable insights for transfer decision making and strategic planning. The models help clubs to make better recruitment decisions, plan for the long term, optimize contract negotiations and manage financial resources more effectively. By reducing financial risks due to objective data, they contribute to both on-field success and long-term financial sustainability. This can thereby indirectly influence a football club's success over the long term.

Keywords: Prediction model, Machine learning, Algorithms, Football player valuation, Feature

Table of contents

1. Introduction

Football stands as the world's most popular sport in terms of both participant and spectator engagement (Cotta et al., 2016). The latest data known about football finances state the revenue generated by professional European football clubs alone is amounted to €29.5 billion[1] in the last season. This underscores the significant economic influence of football (Vroonen et al., 2017). The sport has evolved into a vital contributor to the global economy (Asif et al., 2016).

1.1 Research context

Over the past few decades, there has been a substantial growth in demand for football talent, resulting in astronomous transfer fees amongst individual players. From a managerial standpoint, the pivotal decision confronting football clubs revolving around player transfers has massive influence on a club's prospects for long-term success (Pawlowski et al., 2010). Szymanski and Smith (1997) developed a model showing a linear relationship between profit margins and league performance with revenue influenced by league position. Subsequent studies by Sakınç et al. (2017) confirmed that league position significantly drives revenues for professional football clubs and vice versa. Since transfer fees of players have direct impact on revenue (Supino & Marano, 2024), transfer decision making underlines the importance for professional football clubs.

Despite the significant financial implications and the critical role of player transfers in the success of football clubs, current player valuation models are limited in scope and accuracy. Existing models mainly focus on a narrow subset of features such as player characteristics, performance metrics, crowd-judgement, player potential, or team attributes and often only integrate a few of these aspects. (Behravan & Razavi, 2021; Felipe et al., 2020; He et al., 2015; Herm et al., 2014; Lee et al., 2022; Müller et al., 2017; Yiğit et al., 2020). This fragmented approach results in models that may not fully capture the complexity of factors influencing player market values. Moreover, many of these models are based on outdated methodologies, failing to enhance predictive accuracy through the use of more relevant data and newer methodologies. This gap results in a lack of understanding performance of the latest machine learning techniques for predicting player valuations, features that influence the valuation and its corresponding practical use for football clubs.

Therefore, the primary problem this research seeks to address is the development of a more comprehensive and accurate player valuation model. This model will integrate features from all relevant subsets and apply the latest machine learning methods, structured specifically by player positions, to provide critical knowledge and understanding of the performance of the latest machine learning techniques for predicting player valuations. By addressing this gap, the research aims to improve the reliability of player valuations, ultimately extending existing knowledge about the latest machine learning methods and their corresponding performance on player valuation. It also provides football clubs with more precise and actionable insights for transfer decision making and strategic planning. This can thereby indirectly influence a football club's success over the long term (Pawlowski et al., 2010).

[1] Annual Review of Football Finance 2023 statesthat the revenue generated by European football clubs is amounted to €29.5 billion (Deloitte, 2023)

2. Theoretical framework

According to Kumar (2013), the integration of machine learning is used to revolutionize football analytics, uncovering insights that elude human analysis. Football analysis software allows you to combine multiple tracking and event data streams with custom parameters to create AI-driven actionable insights[2]. This suggests that football analytics, supported by machine learning, will find widespread application in areas such as player valuation facilitated by the increasing availability of relevant data. However, football is a challenging game to analyze due to its underlying nature. The continuous opposition and dynamic structure, in combination with the tactical aspect, make invasion games like football more complex than other game forms (Kumar, 2013). Thus, more difficult to analyze adequately (Kumar, 2013). Any performance analysis with data in invasion games should therefore be structured by the help of a notational analysis system (Pollard et al., 2013; Tenga, 2010).

Historically, player valuation has been a challenge in the football industry. Collaborating with multiple universities around the world, researchers utilized algorithms to objectively determine the value of each player, solely based on their performance data rather than subjective opinions of experts and fans (Müller et al., 2017). Similarly, KPMG, in partnership with OptaSports, a prominent football analytics company, developed a benchmarking model for player valuation[3]. These initiatives underscore the growing interest in the objective to use standardized metrics to evaluate players. Thereby, aiming to mitigate the risks of overvaluation or undervaluation often driven by subjective assessments.

2.1 Distinguishment between player market value and transfer fee

In existing literature there is a distinguishment between a player's market value and a player's respective transfer fee. A player's market value refers to the estimated worth at which a team could sell the player's contract to another team (Herm et al., 2014). Unlike transfer fees which reflect the actual prices paid in the market, market values serve as estimations of these fees and thereby playing a crucial role in transfer negotiations (Müller et al., 2017). In this study, the focus is on a player's market value since it is the most profound way when comparing the model to actual transfers from the past.

2.2 Features that influence the market value of football players

When exploring features that influence a player's market value, it can be concluded that there are many different types of features. When delving into relevant literature, it can be concluded that there are many features, each containing its own level of impact on the valuation of a player, with one being more significant than the other. Overall, these features can be divided into five subsets, namely: *player characteristics, player performance, player popularity, player potential* and *team features*. A description of the most influential features, known from pre-assessing existing literature, is provided onwards.

Player characteristics

Player characteristics consist of both physical and demographic traits, with *age* being a key determinant of market value due to its reflection of experience and ability (Carmichael & Thomas, 1993).

[2] Function of data for football analyses. https://www.scisports.com/services/performance-analysis/

[3] KPMG football benchmark. https://www.footballbenchmark.com/methodology/player_valuation

Research indicates that players typically see a rise in their value until their mid-twenties, followed by a decline (Bryson et al., 2013). Additionally, *height* has been identified as a significant factor in determining salary returns, as it correlates with strong heading ability. This can impact goal-scoring or goal prevention (Fry et al., 2014). Studies have also explored the impact of *footedness* on player valuation, with findings suggesting that ambidextrous players command higher salaries (Bryson et al., 2013) and this has impact on a player's market value (Herm et al., 2014). *Nationality* is another important factor (Frick, 2007), with research suggesting biases in valuation based on players' origins (Garcia-del-Barrio & Pujol, 2007). Furthermore, *player position* plays a crucial role in estimating market value, with salaries and transfer fees varying based on performance and popularity (He et al., 2015; Müller et al., 2017). He et al. (2015) state that attackers tend to receive greater attention and rewards compared to goalkeepers due to their visibility on the field and their capacity to attract crowds. At last, a player's *mentality* is a key aspect as well (Yiğit et al., 2020).

Player performance

Player performance is assessed using various metrics to gauge their market value. *Goals*, encompassing field goals, headers and penalties serve as a measure of scoring ability and are thus pivotal in determining performance (Carmichael & Thomas, 1993). In addition to goals, *assists* are a key factor for measuring performance (Müller et al., 2017). Furthermore, researchers often analyze other performance indicators to in order to determine value and transfer fees. *Passing accuracy* is a commonly utilized metric (Herm et al., 2014) along with statistics on *duels* (such as tackles in the form of clearances) (Inan & Cavas, 2021), *dribbling* success rates (Medcalfe, 2008), *fouls committed* (He et al., 2015), and disciplinary actions like *yellow and red cards (Kiefer, 2012)*.

Crowd-judgement

In football, *crowd-judgement* also plays a significant role regarding market value (Franck & Nüesch, 2012; Müller et al., 2017). The demand for football players is often influenced by their ability to attract crowds, regardless of their on-field performance (Franck & Nüesch, 2012). A player's off-field image impacts merchandise sales and earnings from image rights, leading scientists to explore popularity-related factors in the football transfer market (Hofmann et al., 2021). Popular athletes possess commercial value which is beneficial for their clubs (Arai et al., 2014). Transfermarkt for instance is a leading website for football transfer market data, determined through *crowd-judgement*. Members of the site propose and discuss player market values, which are then aggregated to form final estimates. This method leverages the "wisdom of crowds" concept (Surowiecki, 2005), suggesting that collective judgment can be as accurate as expert opinions.

Player potential

The last significant subset is the *potential* of a player. Al-Asadi and Tasdemır (2022) found that the potential of a player had the highest correlation with the value of that player. The potential of a player is calculated by adding the player's *age*, *international reputation*, and the player's actual game history (performance history) of the player to the overall rating score (Lee et al., 2022). Therefore, *potential* is always equal to or higher than the overall rating.

Team features

Felipe et al. (2020) conducted research in the most influential features and impact of

team features on market value. From extensive studying and testing, *team level*, *birth month*, *league*, *place of play*, and *player's age* influence the players' market values most (Felipe et al., 2020).

2.3 Data sources for feature collection

In order to build a valid model to assess the market value of football players, valid and reliable data sources are obligatory. In existing literature, many scientists rely on data from the Football Manager simulation game renowned for its advanced and detailed database (Yiğit et al., 2020). Football Manager's partnership with Prozone enables its data integration into Prozone Recruiter, utilized by top clubs for player scouting. This data is being revised annually by over 1000 professional scouts worldwide (Yiğit et al., 2020). Football Manager's database includes 49 individual player attributes categorized into technical, mental, physical, and goalkeeping abilities, each rated on a scale of 0 to 20. It offers a unique perspective compared to traditional and volatile in-field statistics. Football Manager's data fully capture a player's performance and is evaluated by experts[4]. Therefore, it provides a comprehensive assessment considering various environmental factors.

Another data source for assessing the market value is SOFIFA or EA24, formerly known as the videogame FIFA. A big portion of researchers acquires extensive and reliable football data, due to its ability in predicting match outcomes with success (Prasetio, 2016). Prasetio (2016) claims that SOFIFA deemed comparable or superior to other football data sources. The EA Sports video game series provides detailed information on European football players covering physical, mental, and technical skills. All the data is accessible on the official website and through the game itself. EA Sports employs real-life scouts to assess player skills which influence in-game ratings, with over 300 fields and 35 attributes determining player ratings (Max 100). Despite lacking a scientific formula for determining market value, SOFIFA ratings are relied upon by scouts, potentially introducing biases.

Furthermore, to acquire remaining values and features, many researchers utilize Transfermarkt due to its wide range of data available (Müller et al., 2017). The domain utilizes its own value prediction model based on crowd-judgement data and aggregated individual estimates (Müller et al., 2017). Studies have shown that Transfermarkt's market values correlate well with expert estimates and player salaries (Bryson et al., 2012; Franck & Nüesch, 2011; Torgler & Schmidt, 2007), making it a valuable resource for research and media.

At last, WhoScored is widely used in the literature to gather required data. This data is containing information about players match records and their respective club's match records. A player's match records consist of features that highlight a player's performance in a season. The club's match record contains information about the performance of the club in that particular season.

2.4 Description of previous created player valuation models through machine learning

To gain a comprehensive overview, related works and their limitations are described at first. In appendix one, a table is given to gain clear insights in the most relevant works for this study.

[4] Football Manager 2024. Sports Interactive 2024

Lee et al. (2022) enhance market value prediction using an optimized LightGBM model with hyperparameter tuning via the Tree-structured Parzen Estimator (TPE) algorithm. Feature importance is determined through the SHapley Additive exPlanations (SHAP) algorithm. Compared to baseline regression models and gradient boosting models without hyperparameter optimization, their optimized LightGBM model achieves significantly higher accuracy. This approach enhances prediction accuracy and provides interpretability. In order to further improve the model's performance, the researchers could have made use of another potent optimized ensemble model such as XGboost or Catboost in combination with employing TPE Bayesian optimization methodology which was not used in this study. Following this optimization step, the researchers could implement the stacking ensemble technique. This is a meta-learning-based ensemble approach that learns to effectively combine multiple models, to combine the optimized GBM, LightGBM, XGboost, and Catboost models. This comprehensive approach is used to deliver an advanced ensemble model for prediction, characterized by better efficiency and performance in the domain of sports analytics.

Al-Asadi and Tasdemır (2022) proposed a quantitative approach using machine learning algorithms applied to FIFA 20 player performance data sourced from sofifa.com. Four regression models, linear, multiple linear, decision trees and random forests were employed to estimate market values and identify key determining factors. Results demonstrate the superiority of random forests in accuracy and error reduction. This objective method offers efficiency and improved performance compared to prior works. The researchers furthermore, suggest exploring additional features to enhance prediction accuracy in potential future research.

Behravan and Razavi (2021) built a machine learning model using the FIFA 20 dataset. In their study, they used hybrid regression. This is a combination of particle swarm optimization (PSO) and support vector regression SVR). According to the authors, the RMSE and MAE for their method are 2,819,286 and 711,029.413, respectively. These results indicate that their method has a significant advantage over other methods of estimating the market value of football players. The study proposes several avenues for further research to enhance the accuracy and effectiveness of player value estimation models in football analytics. Firstly, there's a suggestion to explore additional player attributes beyond the 49 considered in the current study, which could boost the model's accuracy and robustness. Secondly, the integration of advanced optimization techniques, beyond the particle swarm optimization (PSO) and support vector regression (SVR) employed thus far, is recommended to refine feature selection and parameter tuning. This can potentially lead to more precise estimations. Lastly, there's a suggestion to explore methods for enhancing the interpretability and explainability of the estimation model, offering better insights into the factors influencing players' market values.

Felipe et al. (2020) investigated the impact of team features and player positions on market value. Their regression analysis highlighted team level, birth month, league, position, and player age as most influential factors. Attacking midfielders born in the first quarter were particularly valuable. The researchers outline future research avenues to explore the determinants influencing the market value of professional footballers in Europe.

Firstly, they suggest investigating additional factors beyond those considered in the current study, such as performance indicators and popularity to further analyze their impact on player market value. Secondly, the authors propose examining the effects of various independent features including age, round achieved, previous transfers, and minutes played on perceived market value. These future research directions aim to deepen understanding of the complex factors shaping player market values.

Müller et al. (2017) introduced a multi-level regression technique for market value estimation, leveraging a dataset containing player characteristics, performance metrics, and popularity indicators. Their model was significantly more accurate for low- to medium-priced players, whereas the crowd tend to be more accurate for high-priced players. Future research avenues for this study contains investigating the effectiveness of data analytics in scouting young and lesser-known players, particularly in minor leagues where player visibility may be lower. Furthermore, the authors advocate for the incorporation of additional indicators of market value, including league-level, club-level, and individual-level features to enhance the accuracy and robustness of estimation models. Considering sentiment analysis of social media data alongside volume metrics is also suggested as a potential avenue for improving predictive models. Lastly, the analysis of minor league data is proposed to broaden the scope of research and provide insights into market dynamics across different levels of competition.

Herm et al. (2014) introduced a method to estimate transfer fees based on five talent features, highlighting age's inverse correlation with market value. However, reliance on community evaluations introduces potential biases or knowledge gaps, posing a limitation. Future research opportunities include exploring additional features impacting market value and analyzing actual community discussions to compare evaluation processes and effectiveness.

Franck and Nüesch (2012) explored the impact of player talent and popularity on market value, measuring talent across twenty criteria. Using an OLS regression model, they concluded that player popularity positively influences market value. Further research of factors determining the superstar theory is suggested to enhance generalizability of results.

Stanojevic and Gyarmati (2016) proposed a methodology to estimate market values based on player performance data, constructing multiple models using supervised learning and data from Transfermarkt and InStat. The models, built on 45 predictors, outperformed market value estimates from Transfermarkt in relation to team performance. A limitation of this study is the use of older supervised learning methods.

He et al. (2015) developed a model to economically assess all La Liga players, with potential application to other leagues. They attempted modelling the performance over the entire set of players, but failed to find satisfactory results. After focusing on the forward players specifically, it became possible to model their performance. The primary limitation of this study lies in its reliance on community evaluations, which may be subject to bias or limited knowledge. Furthermore, the model could be applied to other leagues and be extended to other player positions to create a more accurate estimation.

Yiğit et al. (2020) aimed to establish a football player value assessment model using machine learning techniques. The proposed models were primarily based on intrinsic features of individual players sourced from the Football Manager video game. To accomplish this, different value assessment models were conducted using advanced supervised learning techniques like ridge and lasso regressions, random forests, and extreme gradient boosting. The actual transfer values of players have found to be closer to their model's valuations after comparison with other models. The study suggests potential enhancements through more advanced techniques like deep learning, particularly artificial neural networks, during ensemble and inflation steps. Additionally, utilizing the lightGBM technique, which is increasingly popular in data science, could further improve the model.

Majewski (2016) delved into the factors influencing the valuation of forward players, aiming to pinpoint the most influential aspects. Analyzing data from 150 renowned attackers sourced from Transfemarkt, the researcher employed the generalized least squares method to identify significant factors. His findings underscored the impact of goals, assists, team value, and FIFA rating points on the market value of attacking players. While acknowledging the study's focus on forwards as a strength, its exclusive emphasis on this position may be considered a limitation, so expanding the model to other positions is recommended.

2.5 Research gap and problem statement

After investigating previous conducted research around player valuation, it became clear that determinants influencing player values in football is gaining popularity in the field of research due to its influence on club's success (Franceschi et al., 2023). Interestingly, existing player valuation models are based on different (sometimes outdated) subsets of features. A subset is a set of features that influence the market value of a player. Subsets of the latest models are either reliant on *player characteristics*, *performance*, *crowd-judgement*, *player potential* or *team features* or there is a combination of a few subsets (Behravan & Razavi, 2021; Felipe et al., 2020; He et al., 2015; Herm et al., 2014; Lee et al., 2022; Müller et al., 2017; Yiğit et al., 2020). Yet, there is no model created based on features integrated from all the different subsets. Thus, there is a lack of insight in models based on combinations of features from all the different subsets together. There is also insufficient insight in models created with the latest available machine learning methods. It is important to address these gaps because it leads to more critical knowledge and understanding of performance of the latest machine learning techniques for predicting player valuations through a more comprehensive approach.

2.6 Research objective

The main purpose of this study is to provide this field of research with more precise insights into the performance of the latest algorithms regarding player valuation predictions. The theoretical contribution of this study, and primary objective, is to address the existing gaps in player valuation models within football. While previous models have focused on features from subsets like *player characteristics*, *performance*, *popularity*, *player potential* or *team features,* this study aims to integrate features from all these subsets into a comprehensive model. The model will be employed through the latest machine learning methods and will be

based on player positions (attackers, midfielders, defenders and goalkeepers). By doing so, it provides a better understanding of the performance of the latest machine learning models and their respective player valuations in football. Additionally, clubs can use it as an enhanced and more precise guideline for transfer decision making and team management, thereby indirectly influencing a football clubs success over the long term (Pawlowski et al., 2010).

This study contributes to advancing the field of football analytics by proposing a more comprehensive and accurate approach to player valuation, with practical implications for football clubs. The following research question is created in order to achieve the research objectives:

RQ: *Do the best performing position-based player valuation models, built with the newest machine learning methods, outperform existing player valuation models?*

Phase 1
The first phase is dedicated to identifying new combinations of features from the different subsets of data through the use of feature analyses and machine learning methods. Next, the subsets will be tested based on performance metrics and compared to single subsets from previous research. The corresponding hypothesis for the first phase, based on suggestions and previous research is as follows:

H1: *Features of all subsets together provide for better performance compared to features from single subsets.*

Phase 2
The second phase is dedicated to achieving a more accurate model through the use of different (newer) machine learning methods and test their performance. Therefore, the corresponding hypothesis is constructed:

H2: *The newer algorithms used in this study, combined with ensemble stacking, demonstrate better performance metrics compared to existing player valuation models.*

Based on limitations and future research opportunities from previous research, the model will be built based on player positions to even further enhance its performance. The corresponding hypothesis for that part the second phase, is as follows:

H3: *Position-based player valuation models built in this study, provide better performance metrics compared to existing player valuation models.*

3. Research strategy

In order to achieve satisfying results, this research is conducted in two phases. Several steps will be carried out within these two phases. Throughout the study, the R software package is used to analyze the data since it provides statistical and graphical methods (Lee et al., 2022).

3.1 Phase 1: Identifying influence and new combinations of features

As mentioned earlier, the first phase is dedicated to identifying new combinations of features from the different subsets of data.

3.1.1 Step 1: Dataset with features affecting a player's market value

In order to identify the influence and possible new combinations for a comprehensive position-based model, a dataset is created with market values from different sources (SOFIFA and

Transfermarkt). An average of those values will be used as the dependent variable. Combining values from both SOFIFA and Transfermarkt provides a more comprehensive and reliable dataset. SOFIFA offers values based on detailed player statistics and ratings, while Transfermarkt provides values based on crowd-judgement and real-world transfer data. This combination ensures a robust and well-rounded evaluation of player values. Additionally, the risk of outliers and abnormalities will be mitigated. This approach enhances the accuracy and reliability of the player valuations which provides a more balanced and realistic prediction. Furthermore, averaging the values helps to reduce the influence of the individual biases which hopefully results in a more neutral and fair player value prediction. At last, features from different subsets will be added from WhoScored. These features are identified through research on existing literature.

3.1.2 Step 2: Preprocessing and feature extraction

The next step in this phase is preprocessing and feature extraction of the data in the set. Preprocessing is a crucial step in data mining. It involves the preparation and transformation of data to make it suitable for analysis (Yiğit et al., 2020). This process includes various techniques such as data cleaning, transformation and reduction. In this study, data cleaning is prioritized to ensure accurate results in the models that will be built later on. The steps involved in data cleaning are: removing redundant columns, handling missing values, converting categorical features to numeric values, grouping unique locations into broader categories, converting numeric columns to appropriate data types and scaling features. Additionally, to ensure a good distribution along the dataset, logarithmic transformation will be applied.

3.1.3 Step 3: Dataset description

Step three in this study is about providing a comprehensive description of the final dataset and its features. Additionally, the results of the preprocessing and feature extraction methods are presented. At last, figures and tables are provided to create a clear overview of the dataset.

3.1.4 Step 4: Analysis of features

In order to identify new combinations of features, the features are analyzed based on their influence on the market value of a player. Therefore, the next step in this phase is to first determine the correlation of these independent features with the dependent feature (a player's market value). To achieve this, Pearson's R correlation coefficient is calculated. Features that exceed a correlation coefficient of 0.3 or higher are considered to have a significant relationship with the dependent feature. Once a plot of correlated features is constructed, the next step is to calculate the feature importance of these features on a player's market value. This will be achieved using the SHapley Additive exPlanations (SHAP) algorithm. The SHAP algorithm assigns an importance value to each feature based on its contribution to the model's output (Wang et al., 2024). SHAP helps identify the most influential features in determining player market value by calculating the impact of each feature on predictions. In general, higher absolute SHAP values indicate greater importance of a feature in predicting the target feature (Wang et al., 2024). A correlation analysis is made to ensure a feature selection based on significant influence. In the second phase of this study, feature selection and feature importance are re-assessed through the use of different machine learning algorithms. These algorithms are capable

of optimized feature selection, feature importance and model creation.

3.2 Phase 2: Creating improved models for player valuation

The second phase of this study is dedicated to achieving superior models through the use of different (newer) machine learning methods and build it based on player positions.

3.2.1 Step 1: Machine learning methods for data modelling

The focus in this study is on the ten most promising and performing methods, based on performance of the models from previous research (Behravan & Razavi, 2021; He et al., 2015; Lee et al., 2022; Majewski, 2016; Müller et al., 2017; Yiğit et al., 2020) and their respective limitations. In particular, the methods used to further enhance the models in this research are: *Linear regression, Ridge regression, Lasso regression, Principal Component regression, Partial Least Squares regression, Particle Swarm Optimization* in combination with *Support Vector Regression, LightGBM* in combination with *Bayesian optimization*, *XGBoost* in combination with *Bayesian optimization, CatBoost* in combination with *Bayesian optimization* and a *Meta-model.* A complete description of all the methods is provided in chapter five of this study.

3.2.2 Step 2: Feature selection

For each position-based model a feature selection will be made based on the influence of features. Some of the algorithms are capable of optimized feature selection, feature importance and model creation without having to use extra analysis of features as mentioned earlier. Logically, those are the features that are used for these models. For the other models (*Linear regression, Ridge regression, Lasso regression, Principal Component regression and Partial Least Squares regression*) feature selection is based the analysis from phase one of this research.

3.2.3 Step 3: Splitting dataset in train-set and test-set

After completing the data-preprocessing stage and defining the subsets of relevant features, the dataset is being divided into four sets that are based on position lines. The next step is to randomly assign 80% of the data per set for training the classifiers and reserving 20% for testing purposes.

3.2.4 Step 4: Model evaluation and validation

In order to validate the training methods, k-fold cross-validation is used where k = 10. In addition, the datasets will be trained 100 times. Each machine learning model aims to address a particular problem using diverse datasets. In regression problems, common evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the coefficient of determination (R2) (Lee et al., 2022). These metrics help assess the performance of the model in accurately predicting outcomes. The metrics to evaluate and validate the models in this study are: *RMSE and R2.*

3.2.5 Step 5: Conclusion based on findings

The last step of this study implies writing a conclusion and discussion based on findings from the feature analyses and the performance metrics of the created models. Furthermore, possible limitations and future research avenues are described. Finally, an examination of the possible hypotheses is provided with an answer to the research question to ensure comprehensive coverage of the study's objectives.

4. Description of dataset

The dataset used in this study integrates comprehensive data of football players that is gathered from SOFIFA, Transfermarkt, and WhoScored. It covers data from the 2023/2024 Eredivisie season. The data includes a wide range of features categorized into monetary value, player characteristics, player performance, player potential, team features, and crowd-judgement ratings. It covers 99 features in total. A complete overview of the features with explanation can be found in appendix 2.

Monetary value features

Monetary value features include the market values of players as calculated by Transfermarkt and SOFIFA. Additionally, the mean player values are integrated. The mean value of a player is the dependent variable in this study. Furthermore, weekly salaries and release clauses are gathered. These provide insights into the market worth of players (Lee et al., 2022). All monetary values are treated as ratio data with no predefined limitations on their range. This reflects the real-world variability in player market values.

Player characteristics

Player characteristics encompass features that define a player's personal and professional profile (Carmichael & Thomas, 1993). This includes demographic information such as nationality, age, date of birth, height and weight. Additionally, it captures the player's football-specific features like club affiliation, contract details, overall ratings, positions and position-specific ratings. Features like preferred foot, weak foot rating, skill moves, attacking and defensive work rates are also included. These data are primarily scraped from SOFIFA, ensuring a detailed and accurate representation of each player's profile. The characteristics data are a mix of nominal, ordinal, interval and ratio data types. This reflects the diverse nature of player attributes. A description of the ability feature calculations can be found in appendix 3.

Player performance

Performance metrics are essential for evaluating a player's contribution on the field (Carmichael & Thomas, 1993; He et al., 2015). This dataset includes detailed statistics on player appearances, minutes played, goals, assists, yellow and red cards, shots per game, passing accuracy, aerial duels won, key passes, dribbles, tackles, interceptions and other performance-related features. These performance indicators are gathered from WhoScored, providing a comprehensive view of each player's on-field performance during the 2023/2024 season. Most of these features are treated as ratio data.

Player potential

Player potential is a predictive measure of a player's future performance, based on their current abilities, age and international reputation (Al-Asadi & Tasdemır, 2022). This feature is important for forecasting a player's career trajectory and market value growth. They are scraped from SOFIFA and are treated as interval data, ranging from 1 to 99.

Team features

Team performance attributes provide context to individual player metrics by reflecting the overall success and standing of their related clubs (Felipe et al., 2020). This includes team-level statistics such as total goals scored, conceded, goal difference, victory points and the number of wins, draws and losses. Team standings at the end of the season are also included. These features are scraped from WhoScored and are primarily ratio data. Team standings are being treated as

ordinal. They are critical for understanding the team dynamics and the environment in which players operate (Felipe et al., 2020).

Crowd-judgement ratings

Crowd-judgement ratings offer a qualitative assessment of player performance, aggregated from crowd-sourced evaluations and expert opinions (Franck & Nüesch, 2012; Müller et al., 2017). These ratings range from 0 to 10 and provide an additional layer of analysis. It thereby reflects public and expert perceptions of player performance (Franck & Nüesch, 2012; Hofmann et al., 2021; Müller et al., 2017). This feature is scraped from WhoScored and treated as interval data.

4.1 Data preprocessing and feature extraction methods

The initial step in preparing the dataset involves merging the diverse sources of collected data into a single comprehensive dataset. The first step is to inspect the initial dataset for any abnormalities. The distribution plot and boxplot of figure 1 and 2 show that the initial dataset is skewed to the right. This indicates that player values are not normally distributed.

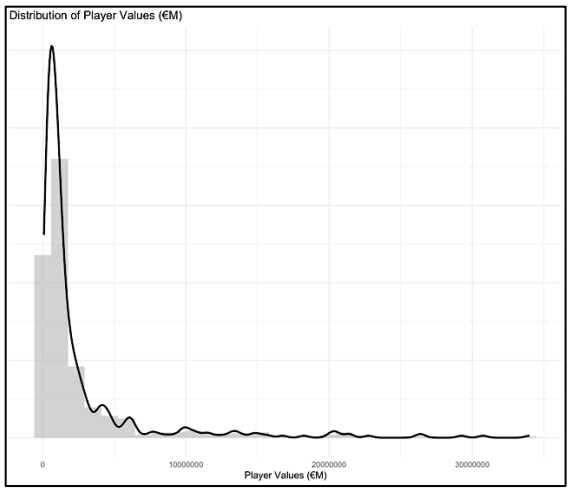

Figure 1: Distribution of player values before logarithmic transformation

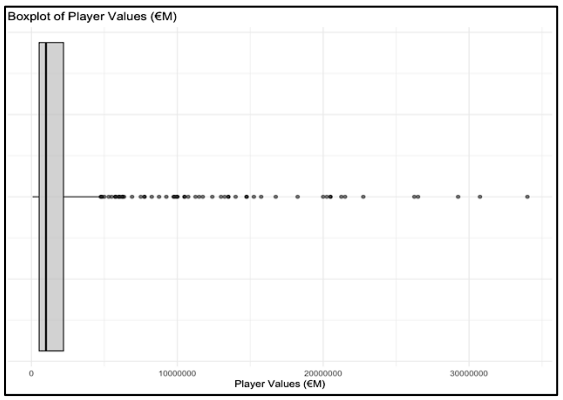

Figure 2: Boxplot of player values before logarithmic transformation

After inspecting the dataset, several steps need to be carried out to ensure the dataset's integrity and usability. The first step is adressing missing data by performing listwise deletion. Players with any missing values across the columns are removed from the dataset which ensures a complete dataset for subsequent analyses. This step is crucial to avoid biases and inaccuracies that could arise from building models upon incomplete data. The following step refined the dataset by removing unnecessary columns that don't contribute to a player's value.

All variables are converted into numeric values to facilitate mathematical and statistical analysis for the regression models. Converting the categorial variables is achieved by assigning a unique numeric code to each category. This allows for the inclusion of these variables in the modeling process.

Following the conversion to numeric values, the dataset underwent a scaling process. Scaling is necessary to standardize the range of the variables which ensures that each feature contributed equally to the analysis. Standardization is performed using z-score normalization which adjusts the data to have a mean of zero and a standard deviation of one.

Logarithmic transformation is applied to address issues of skewness and non-normality in the data distribution. This transformation is particularly effective for data that contains a right-skewed distribution. By applying the transformation, the data was normalized which results in a more symmetric distribution with reduced skewness. Figure 3 and 4 show the results of this process.

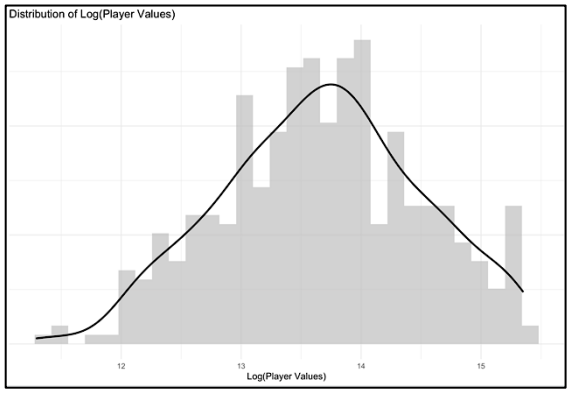

Figure 3: Distribution of player values after logarithmic transformation

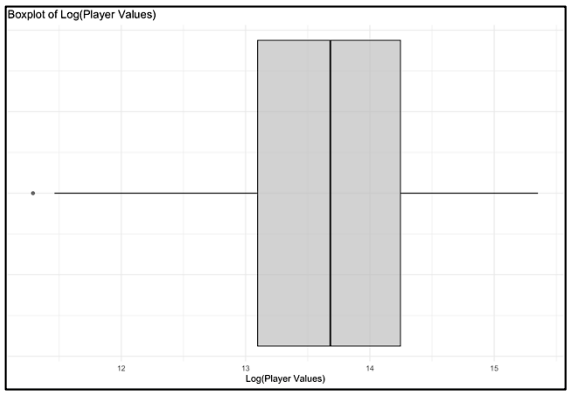

Figure 4: Boxplot of player values after logarithmic transformation

Overall, the dataset is now more equally distributed. The data is now more suitable for building the prediction models even though it shows a light skewness to the left.

5. Description of machine learning methods

Each of the methods and techniques possess unique strengths that can be leveraged depending on the specific requirements of the football player valuation models. Linear regression offers simplicity and interpretability (Al-Asadi & Tasdemır, 2022; Puccio, 1999; Wu et al., 2008), while PSO combined with SVR provides a robust optimization approach for (non-)linear relationships (Behravan & Razavi, 2021). LightGBM, XGBoost and CatBoost in combination with Bayesian optimization, offer state-of-the-art performance for complex datasets. This makes them suitable for predictive modeling (Lee et al., 2022; Yiğit et al., 2020). SHAP values add an extra layer of trust, validation and interpretability regarding the feature selection of the models (Lee et al., 2022). A detailed description for every method that is used in this study is given below.

5.1 Linear regression

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables (Montgomery et al., 2021; Wu et al., 2008). It operates by fitting a linear equation to the observed data, where the coefficients of the equation represent the extent of influence that each predictor variable has on the dependent variable. The primary goal of linear regression is to find the best fitting straight line through the data points. The line minimizes the sum of the squared differences between the observed values and the values predicted by the model. This method assumes a linear relationship between the predictors and the outcome. This assumption makes it straightforward to interpret the results. However, its simplicity can also be a limitation when dealing with complex, non-linear relationships for player valuation prediction (Al-Asadi & Tasdemır, 2022).

5.2 Ridge regression

Ridge regression is a linear regression technique that addresses multicollinearity

by adding a regularization term to the least squares cost function (Hoerl & Kennard, 1970). This regularization term is the squared magnitude of the coefficients, which penalizes large coefficients and shrinks them towards zero. This is stabilizing the estimates and improves the model's generalizability. By controlling the complexity of the model, Ridge regression balances the trade-off between bias and variance. Hoerl and Kennard (1970) state that this leads to more reliable predictions in the presence of highly correlated predictors.

5.3 Lasso regression

Lasso regression (Least Absolute Shrinkage and Selection Operator) is a type of linear regression that performs both variable selection and regularization to enhance prediction accuracy and interpretability (Tibshirani, 1996). By adding a penalty equal to the absolute value of the coefficients to the cost function, Lasso regression forces some of the coefficient estimates to be exactly zero. This effectively reduces the number of predictors in the model. It makes Lasso particularly useful for models with a large number of features as it simplifies the model by selecting only the most relevant variables (Tibshirani, 1996).

5.4 Principal Component Regression

Principal Component Regression (PCR) combines Principal Component Analysis (PCA) with linear regression to address multicollinearity and reduce the dimensionality of the data (Jolliffe, 1982). PCA transforms the original predictors into a smaller set of uncorrelated components. This captures the maximum variance in the data. These principal components are then used as predictors in a linear regression model. By focusing on the most important components, PCR reduces overfitting and improves the model's predictive performance. The algorithm particularly help in cases where the original predictors are highly correlated (Jolliffe, 1982).

5.5 Partial Least Squares regression

Partial Least Squares (PLS) regression is a statistical method that models the relationship between input features and the target variable. This is achieved by extracting latent variables that maximize the covariance between the predictors and the response (Wold et al., 1984). Unlike PCR, which only considers the variance in the predictors, PLS also takes into account the response variable. This results in components that are more relevant for prediction. This approach makes PLS Regression particularly effective for datasets with many correlated predictors (Wold et al., 1984).

5.6 Particle Swarm Optimization with Support Vector Regression

Particle Swarm Optimization (PSO) is an optimization technique inspired by the social behavior of birds flocking or fish schooling (Kennedy & Eberhart, 1995). In this method, potential solutions are considered as particles that "fly" through the solution space. They adjust their positions based on their own experience and that of their neighbors to find the optimal solution (Kennedy & Eberhart, 1995). Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) tailored for regression tasks. It aims to find a function that approximates the relationship between input features and the target variable within a specified margin of error (Vapnik & Vapnik, 1998). Combining PSO with SVR allows for the automatic optimization of SVR's hyperparameters. This enhances the performance by navigating the solution space more effectively than traditional grid search methods (Behravan & Razavi, 2021).

5.7 LightGBM with Bayesian Optimization

Light Gradient Boosting Machine (LightGBM) is a highly efficient and fast implementation of gradient boosting. It leverages tree-based learning algorithms (Ke et al., 2017). The algorithm is designed to handle large-scale data with high efficiency by using techniques such as histogram-based decision tree learning and leaf-wise growth. This leads to faster training and reduced memory usage (Ke et al., 2017). Bayesian optimization can be employed to further enhance LightGBM's performance (Bergstra et al., 2011). Bayesian optimization builds a probabilistic model of the objective function and uses this model to select the most promising hyperparameters to evaluate. This process improves the model's performance by focusing the search on the most relevant areas of the hyperparameter space (Lee et al., 2022).

5.8 XGBoost with Bayesian Optimization

Extreme Gradient Boosting (XGBoost) is an optimized distributed gradient boosting library that has gained popularity for its speed and performance in predictive modeling (Chen & Guestrin, 2016). XGBoost incorporates a range of advanced features such as regularization to prevent overfitting and sparsity-aware learning that makes it robust to missing data (Chen & Guestrin, 2016). Similar to LightGBM, XGBoost's performance can be further enhanced using Bayesian optimization (Bergstra et al., 2011). XGBoost's hyperparameters can be fine-tuned more efficiently than traditional methods by employing Bayesian optimization. This ensures that the model achieves high accuracy with optimal computational resources (Bergstra et al., 2011; Lee et al., 2022).

5.9 CatBoost with Bayesian Optimization

Categorical Boosting (CatBoost) is a gradient boosting library specifically designed to handle categorical features without the need for extensive preprocessing (Prokhorenkova et al., 2018). It employs innovative techniques such as ordered boosting and efficient handling of categorical data to improve both speed and prediction accuracy (Prokhorenkova et al., 2018). Bayesian optimization is utilized to optimize CatBoost's performance. This optimization approach aims to build a probabilistic model of the performance of different hyperparameters. Furthermore, it guides the search towards the most promising configurations which improves the model's accuracy and efficiency.

5.10 Meta-model through ensemble stacking

Ensemble stacking is a technique that combines multiple models into one Meta-model. It enhances predictive performance by leveraging the strengths of each individual model (Wolpert, 1992). In this approach, the best-performing models for each player line are selected and stacked together. The meta-model is trained on the outputs of these base models. It learns to optimally combine their predictions which hopefully results in a more accurate and robust model. This stacked approach captures the diverse strengths of each base model which aims to provide superior predictions for each player line (Breiman, 1996).

5.11 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) is a game-theoretic approach to explain the output of machine learning models (Wang et al., 2024). SHAP values assign an importance score to each feature based on its contribution to the model's predictions

(Wang et al., 2024). SHAP helps in understanding which features are most influential in determining the target variable by computing the impact of each feature on the predictions. Higher absolute SHAP values indicate greater importance. This provides a clear indication of feature significance and interaction effects within the model (Lee et al., 2022; Wang et al., 2024).

6. Findings

In the first part of this chapter, an analysis of the feature selection and relation to the dependent variable per line is presented. The results of the prediction models developed to estimate football player values are presented in the second part. The analysis includes ten models and examines the predictive accuracy trough performance metrics. The findings provide insights into the key features influencing player valuations per line and the performance of the models.

6.1 Feature selection and importance per line

For predicting football player values, feature selection and importance are determined using Pearson's correlation analysis, mean SHAP values and importance scores of the features in relation to the dependent variable.

The first analysis of features show that the Lasso regression model is the only model that employs 14 features from every category. All other models use 43 features that also originate from every category. The analysis of feature importance across all models shows that certain categories are consistently more represented and significant for predicting player values across different lines of positions. Player characteristics and player performance are the most represented and important categories across all models, regardless of the specific algorithm that is used. Features like overall rating, potential, wage, minutes played and goals are consistently significant. This indicates their crucial role in predicting football player values. Despite being utilized across all models, crowd-judgment features are less important compared to the other categories. It is interesting to see that all models use features from every category. Yet, none of the algorithms use exactly the same set of features except for the linear regression models (due to the outcome of feature analyses). A complete overview of the most important features and their influence per line of position for all models, can be found in appendices five till ten.

Attackers

Pearson's correlation analysis and importance scores reveal that overall rating and wage from the monetary value category are the highest positively correlated features for attackers. Attributes related to player performance and physical characteristics such as finishing, shooting and stamina also show significant importance. Negative correlations are observed with team features like goals against and team standing. Mean SHAP values emphasize that goal acquisition from the team features category is the most critical feature, followed by wage and overall rating. Player potential, attacking skills and metrics like man of the match awards and minutes played are also influential. Crowd-judgement features play a role although they are less critical than other categories.

Midfielders

Overall rating and wage again demonstrate the highest positive correlations and scores but this time for midfielders. This indicates the importance of features within the monetary value and player characteristics

categories. Skills such as ball control, dribbling, vision and composure are highly significant within the player performance category. Player potential and movement reactions are also essential. Negative correlations are found with team features such as goals against and team standing. Mean SHAP values highlight overall rating and potential as the most significant features. The following significant features are goal acquisition and minutes played. Other relevant features include date of birth and wage.

Defenders

Defensive characteristics such as standing tackle and sliding tackle show the highest positive correlations and scores in relation to player value of defenders. This highlights the importance of the player performance category. Movement reactions, wage and potential are also significant. Physical attributes like jumping and mental attributes such as composure play important roles. All these features originate from the player performance category as well. Features from the team features category such as goals against and team standing show negative correlations. Mean SHAP values underscore the importance of overall rating, potential and goal acquisition. Other critical features include minutes played, movement reactions and defensive skills.

Goalkeepers

Features such as wage, goalkeeping diving and overall rating are key for goalkeepers according to Pearson's correlation analysis and importance scores. This highlights the importance of the monetary value and player performance categories again. Specific goalkeeping skills such as reflexes, handling and positioning also show strong positive correlations. Player potential, minutes played and international reputation are also critical. Team features such as team standing and lose points demonstrate negative correlations. According to mean SHAP values, minutes played, potential and goalkeeping diving are the most significant features. Age, reflexes, wage, positioning and overall rating are also important.

6.2 Performance metrics Linear regression

The Multiple Linear regression model performs the worst for attackers, showing a high RMSE of 13.374 and a low R^2 of 0.239. These metrics indicate poor predictive accuracy and explanation of the variance in the dependent variable. This model achieves better results for midfielders with an RMSE of 0.382 and a much higher R^2 of 0.864. For defenders it also performs well. This is indicated by an RMSE of 0.402 and an R^2 of 0.854. The model shows moderate performance for goalkeepers reflected in an RMSE of 0.706 and an R^2 of 0.719. All the metrics of the model per line can be seen in the table below.

Table 5: Performance metrics for Multiple Linear regression model

Prediction model	Line	RSME	MAE	R2
Mulitiple Linear Regression	Attackers	13.374	10.299	0.239
	Midfielders	0.382	0.274	0.864
	Defenders	0.402	0.315	0.854
	Goalkeepers	0.706	0.608	0.719

6.3 Performance metrics Ridge regression

The Ridge regression model significantly improves accuracy in predicting values for attackers with an RMSE of 0.404 and a high R^2 of 0.912. It maintains good performance for midfielders with an RMSE of 0.398 and an R^2 of 0.853. The model shows slightly lower accuracy compared to other models for defenders with an RMSE of 0.419 and an R^2 of 0.88. It performs strongly for goalkeepers, with an RMSE of 0.378 and an R^2 of 0.884. The results of the Ridge

regression model are presented in table 6 below.

Table 6: Performance metrics for Ridge regression model

Prediction model	Line	RSME	MAE	R2
Ridge Regression	Attackers	0.404	0.284	0.912
	Midfielders	0.398	0.260	0.853
	Defenders	0.419	0.318	0.880
	Goalkeepers	0.378	0.249	0.884

6.4 Performance metrics Lasso regression

The Lasso regression model also achieves great performance for attackers with an RMSE of 0.374 and an R^2 of 0.924. It shows good performance for midfielders as well, achieving an RMSE of 0.397 and an R^2 of 0.854. The model performs similarly to other linear regression models for defenders with an RMSE of 0.427 and an R^2 of 0.876. It is also capable of predicting with high accuracy for goalkeepers achieving an RMSE of 0.351 and an R^2 of 0.9. All the metrics are presented in the table below.

Table 7: Performance metrics for Lasso regression model

Prediction model	Line	RSME	MAE	R2
Lasso Regression	Attackers	0.374	0.315	0.924
	Midfielders	0.397	0.255	0.854
	Defenders	0.427	0.307	0.876
	Goalkeepers	0.351	0.260	0.900

6.5 Performance metrics Principal Component Regression

Table 8 shows that the Principal Component Regression model performs well for attackers with an RMSE of 0.413 and an R^2 of 0.908. The model shows similar performance to the Ridge and Lasso regression models for midfielders, achieving an RMSE of 0.39 and an R^2 of 0.859. The model demonstrates comparable performance for defenders with an RMSE of 0.459 and an R^2 of 0.857. It achieves high accuracy for goalkeepers as well with an RMSE of 0.33 and an R^2 of 0.911.

Table 8: Performance metrics for Principal Component regression model

Prediction model	Line	RSME	MAE	R2
Principal Component Regression	Attackers	0.413	0.279	0.908
	Midfielders	0.390	0.278	0.859
	Defenders	0.459	0.325	0.857
	Goalkeepers	0.330	0.212	0.911

6.6 Performance metrics Partial Least Squares regression

The Partial Least Squares Regression model shows lower performance for attackers compared to other models as can be seen in table 9. The model achieves an RMSE of 0.461 and an R^2 of 0.885. It maintains good performance for midfielders with an RMSE of 0.394 and an R^2 of 0.856. The model has a higher RMSE of 0.488 and an R^2 of 0.838 for defenders. This indicates relatively lower accuracy. It shows moderate performance for goalkeepers with an RMSE of 0.480 and an R^2 of 0.813.

Table 9: Performance metrics for Partial Least Squares regression model

Prediction model	Line	RSME	MAE	R2
Partial Least Squares Regression	Attackers	0.461	0.386	0.885
	Midfielders	0.394	0.262	0.856
	Defenders	0.488	0.355	0.838
	Goalkeepers	0.480	0.379	0.813

6.7 Performance metrics Particle Swarm Optimization with Support Vector Regression

Particle Swarm Optimization combined with Support Vector Regression model demonstrates the best performance for attackers and midfielders. The model achieves an RMSE of 0.29 and an R^2 of 0.954 for attackers, which are the best scores across all models. The RMSE of 0.322 and an R^2 of 0.904 for midfielders is also the best score across all models. For defenders,

a similar level of precision can be seen in table 12 with an RMSE of 0.348 and an R^2 of 0.917. Goalkeepers' values are also predicted well under this model with an RMSE of 0.359 and an R^2 of 0.895. These metrics showcase the effectiveness and accuracy of the PSO-SVR model across all positions and excelling with the best scores for two lines.

Table 12: Performance metrics for Particle Swarm Optimization with Support Vector Regression model

Prediction model	Line	RSME	MAE	R2
Particle Swarm Optimization with Support Vector Regression	Attackers	**0.290**	**0.204**	**0.954**
	Midfielders	**0.322**	**0.205**	**0.904**
	Defenders	0.348	0.253	0.917
	Goalkeepers	0.359	0.249	0.895

6.8 Performance metrics LightGBM with Bayesian Optimization

Table 13 presents the performance metrics of the LightGBM model that is optimized with Bayesian methods. The model shows moderate performance for attackers compared to other models with an RMSE of 0.559 and an R^2 of 0.831. It performs better for midfielders achieving an RMSE of 0.364 and an R^2 of 0.877. The model shows a slightly higher RMSE of 0.47 and an R^2 of 0.849 for defenders. This indicates moderate predictive power when compared to the other models. The model's performance decreases slightly with an RMSE of 0.474 and an R^2 of 0.817 for goalkeepers. This reflects the challenges of accurate predictions in this position.

Table 13: Performance metrics for LightGBM with Bayesian Optimization

Prediction model	Line	RSME	MAE	R2
LightGBM with Bayesian Optimization	Attackers	0.559	0.344	0.831
	Midfielders	0.364	0.226	0.877
	Defenders	0.470	0.358	0.849
	Goalkeepers	0.474	0.348	0.817

6.9 Performance metrics XGBoost with Bayesian Optimization

The XGBoost model which is enhanced through Bayesian Optimization demonstrates strong performance across various player positions. The model achieves an RMSE of 0.45 and an R^2 of 0.89 for attackers. Midfielders' values benefit from a robust prediction with an RMSE of 0.366 and an R^2 of 0.876. The model also performs well for defenders, achieving an RMSE of 0.373 and an R^2 of 0.905. The XGBoost model enhanced through Bayesian Optimization demonstrated the highest accuracy for goalkeepers across all models. It achieves an RMSE of 0.314 and an R^2 of 0.92 which highlights the model's effectiveness in this position. The metrics of this model are presented in the table below.

Table 14: Performance metrics for XGBoost with Bayesian Optimization model

Prediction model	Line	RSME	MAE	R2
XGBoost with Bayesian Optimization	Attackers	0.450	0.344	0.890
	Midfielders	0.366	0.248	0.876
	Defenders	0.373	0.282	0.905
	Goalkeepers	**0.314**	**0.212**	**0.920**

6.10 Performance metrics CatBoost with Bayesian Optimization

The CatBoost model combined with Bayesian Optimization delivers strong performance across all player positions. The model achieves an RMSE of 0.386 and an R^2 of 0.919 for attackers. Value predictions for midfielders are also accurate and explain a larger proportion of the variance. The model achieves an RMSE of 0.335 and an R^2 of 0.896. The CatBoost model combined with Bayesian Optimization performs the best for defenders in comparison to the other models. Defenders' values are well-predicted with an RMSE of 0.341 and an R^2 of 0.921. For goalkeepers, the model maintains solid performance achieving an

RMSE of 0.38 and an R^2 of 0.882. Table 15 shows the versatility and strength of the CatBoost model optimized with Bayesian techniques expressed in its performance metrics.

Table 15: Performance metrics for CatBoost with Bayesian Optimization model

Prediction model	Line	RSME	MAE	R2
CatBoost with Bayesian Optimization	Attackers	0.386	0.284	0.919
	Midfielders	0.335	0.224	0.896
	Defenders	**0.341**	**0.251**	**0.921**
	Goalkeepers	0.380	0.260	0.882

6.11 Performance metrics Meta-model

The Meta-model that is created through ensemble stacking, performed well compared to the other models. It outperformed most of the models with some metrics close to the best performing models per line as can be seen in table 16. For attackers, the model achieves an RMSE of 0.304 and an R^2 of 0.95 which indicates high accuracy. It shows strong results for midfielders as well with an RMSE of 0.33 and an R^2 of 0.899. This showcases the model's reliability in predicting midfielder values. The model also performs well for defenders achieving an RMSE of 0.343 and an R^2 of 0.92 which is close to the best-performing models in this category. The Meta-model's robustness works well for predicting values of goalkeepers with an RMSE of 0.324 and an R^2 of 0.915.

Table 16: Performance metrics for Meta-model

Prediction model	Line	RSME	MAE	R2
Meta-model	Attackers	0.304	0.251	0.950
	Midfielders	0.330	0.207	0.899
	Defenders	0.343	**0.242**	0.920
	Goalkeepers	0.324	0.228	0.915

6.12 Overall comparison between the best performing models

After developing all the models, plotted overviews of their performances are created to analyze their RMSE, MAE and R^2 for comparison. Figure 29 on the next page shows the performance of the models in RMSE. Figure 30 and 31 in appendix 11 show the performance in MAE and R^2. Multiple Linear regression is left out of these figures because of its relatively poor performance compared to the other models, although it is used in the Meta-model. Table 17 on the next page shows all the metrics of all models in one complete overview.

Attackers

Figure 29 and table 17 show that the PSO-SVR model performs best for attackers with a RMSE of 0.29 versus a RMSE of 0.304 for the Meta-model. Furthermore, the R^2 of the PSO-SVR model is slightly higher, achieving 0.954 versus 0.95 for the Meta-model. These metrics show that the performances of these models are equally great in terms of predicting accuracy and explaining a large proportion of the variance. Yet, the PSO-SVR model is slightly more sophisticated.

Midfielders

The PSO-SVR model also performs best for midfielders with a RMSE of 0.322 versus a RMSE of 0.33 for the Meta-model. Also, the PSO-SVR model showed a slightly higher R^2 being 0.904 versus 0.899 of the Meta-model. The minor differences in performance demonstrate that the PSO-SVR model is a tiny bit more accurate in predicting and explaining a larger proportion of the variance.

Defenders

For defenders, the CatBoost model with Bayesian Optimization performs best with an RMSE of 0.341. Nevertheless, the Meta-model and the PSO-SVR model are not far behind achieving an RMSE of 0.343 and 0.348 respectively. When analyzing figure 30 in appendix 11, the R^2 of the Catboost

model with Bayesian Optimization is also a tiny bit better than the Meta-model and PSO-SVR model, being 0.921, 0.920 and 0.917 respectively.

Goalkeepers

Figure 29 and table 17 show that the XGBoost model with Bayesian Optimization performs best for goalkeepers, followed by the Meta-model and the Lasso Regression model. The XGBoost model with Bayesian Optimization achieves an RMSE of 0.314 and an R^2 of 0.92, indicating how well it performs in predicting a goalkeepers' value. The Meta-model and Lasso Regression model follow closely with an RMSE of 0.324, 0.351 and an R^2 of 0.915, 0.9.

Table 17: Performance metrics for all models

Prediction model	Line	RSME	MAE	R2
Mulitiple Linear Regression	Attackers	13.374	10.299	0.239
	Midfielders	0.382	0.274	0.864
	Defenders	0.402	0.315	0.854
	Goalkeepers	0.706	0.608	0.719
Ridge Regression	Attackers	0.404	0.284	0.912
	Midfielders	0.398	0.260	0.853
	Defenders	0.419	0.318	0.880
	Goalkeepers	0.378	0.249	0.884
Lasso Regression	Attackers	0.374	0.315	0.924
	Midfielders	0.397	0.255	0.854
	Defenders	0.427	0.307	0.876
	Goalkeepers	0.351	0.260	0.900
Principal Component Regression	Attackers	0.413	0.279	0.908
	Midfielders	0.390	0.278	0.859
	Defenders	0.459	0.325	0.857
	Goalkeepers	0.330	0.212	0.911
Partial Least Squares Regression	Attackers	0.461	0.386	0.885
	Midfielders	0.394	0.262	0.856
	Defenders	0.488	0.355	0.838
	Goalkeepers	0.480	0.379	0.813
Particle Swarm Optimization with Support Vector Regression	Attackers	**0.290**	**0.204**	**0.954**
	Midfielders	**0.322**	**0.205**	**0.904**
	Defenders	0.348	0.253	0.917
	Goalkeepers	0.359	0.249	0.895
LightGBM with Bayesian Optimization	Attackers	0.559	0.344	0.831
	Midfielders	0.364	0.226	0.877
	Defenders	0.470	0.358	0.849
	Goalkeepers	0.474	0.348	0.817
XGBoost with Bayesian Optimization	Attackers	0.450	0.344	0.890
	Midfielders	0.366	0.248	0.876
	Defenders	0.373	0.282	0.905
	Goalkeepers	**0.314**	**0.212**	**0.920**
CatBoost with Bayesian Optimization	Attackers	0.386	0.284	0.919
	Midfielders	0.335	0.224	0.896
	Defenders	**0.341**	**0.251**	**0.921**
	Goalkeepers	0.380	0.260	0.882
Meta-model	Attackers	0.304	0.251	0.950
	Midfielders	0.330	0.207	0.899
	Defenders	0.343	0.242	0.920
	Goalkeepers	0.324	0.228	0.915

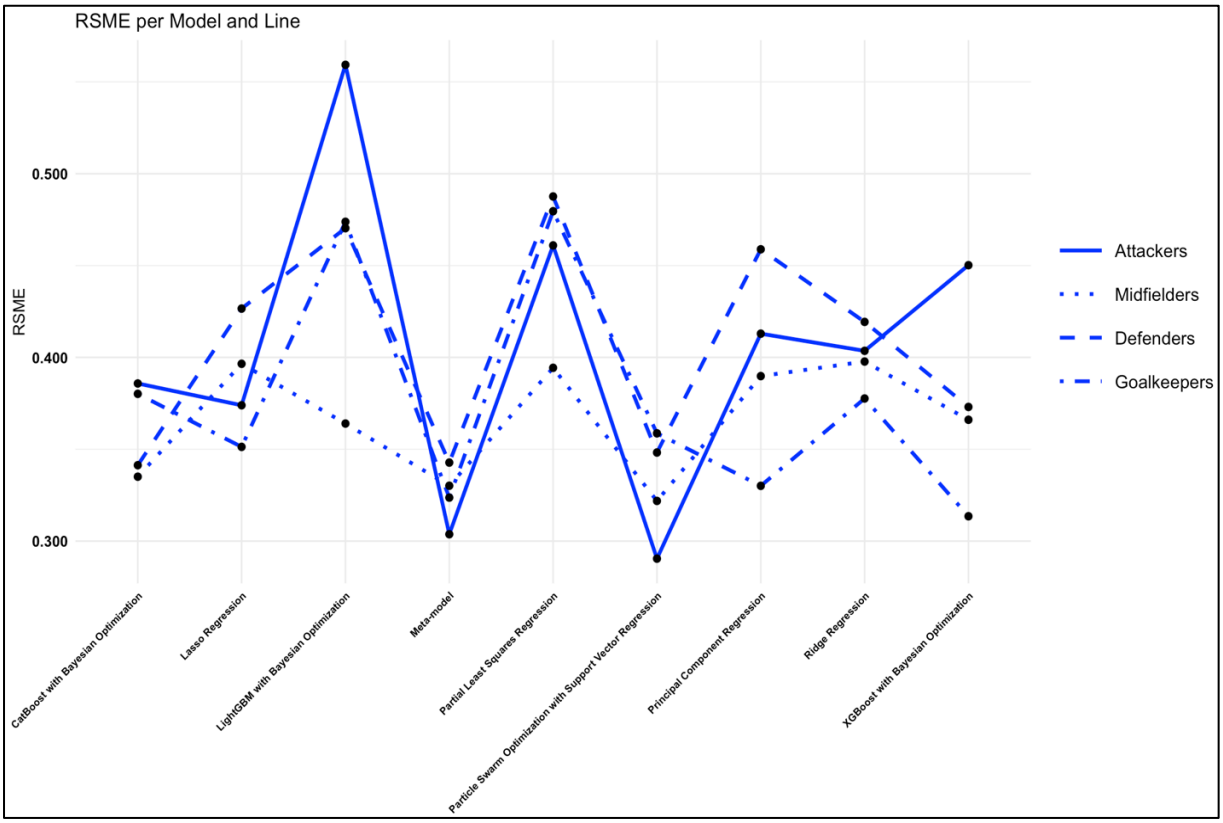

Figure 29: Performance in RMSE per model

Table 18: Performance metrics of models for comparison between studies

Reference	Methods	Line of position	Number of Features	RSME	MAE	R2
Müller et al. (2017)	Multi level regression	all	22	5793.474	3241.733	-
Behravan and Razavi (2020)	Particle Swarm Optimization with Support Vector Regression	all	55	2819.286	711.029	0.74
M. A. Al-Asadi, S. Tasdemir (2022)	Random Forest	all	7	1649.921	576.874	0.95
Lee et al. (2022)	LightGBM with Hyperparameter optimization	all	20	609.42	211.17	-
This study. (2024)	PSO with SVR	Attackers	43	**0.290**	**0.204**	**0.954**
		Midfielders		**0.322**	**0.205**	0.904
	CatBoost with BO	Defenders		**0.341**	**0.251**	0.921
	XGBoost with BO	Goalkeepers		**0.314**	**0.212**	0.920

7. Discussion and conclusion

The primary aim of this study was to evaluate whether position-based player valuation models outperform existing predicting models when they are built with the latest machine learning methods. Therefore, this research was guided by the central question: Do the best-performing position-based player valuation models built with the newest machine learning methods outperform existing player valuation models? The study also explored three sub-questions focusing on the importance of feature selection, the effectiveness of newer algorithms and the benefits of position-based modeling to address the central question.

The analysis of feature selection across all models provides support for the hypothesis. Models that incorporate features from all relevant subsets achieve superior performance compared to using features from a single category. Except for the Lasso regression model which utilized 14 features, all other models leveraged a broader set of 43 features from all categories. This comprehensive approach was crucial for achieving superior predictive accuracy. Features from categories such as: player characteristics and performance consistently emerged as the most Influential across all models. Key features such as overall rating, potential, wage, minutes played and goals were identified as particularly significant in determining player valuations. Interestingly, crowd-judgment features were present in all models but their importance was consistently lower compared to other features. This highlights the reliability of performance-related data over subjective evaluations. The position-specific feature selection underscores the benefit of integrating unique factors relevant to different player roles. It thereby enhances overall model performance.

The findings from this study suggest that Particle Swarm Optimization combined with Support Vector Regression model emerged as the most effective method for predicting valuations of attackers and midfielders. The model exhibited superior predictive accuracy and reliability with the lowest RMSE and the highest R^2 values. The PSO-SVR's performance can be attributed to its robust optimization capabilities, allowing it to navigate complex and non-linear relationships more effectively than traditional models. The Meta-model did not outperform the PSO-SVR even though this was expected. This was likely due to the

constraints of the small dataset which may have limited its ability to leverage the strengths of multiple models simultaneously.

When evaluating model performance for the other positions, the study found that the CatBoost model with Bayesian Optimization performed exceptionally well for defenders. The XGBoost model with Bayesian Optimization showed the highest accuracy for goalkeepers. These findings highlight the importance of selecting machine learning methods tailored to the characteristics and demands of different player positions. Both of the models achieve a much smaller RMSE and a slightly lower R^2 in comparison to previous generated models. These metrics show their superiority. Comparing these results with previous studies presents some challenges, as many existing models use general metrics without accounting for player positions. However, this study's position-specific approach provides a more granular and accurate analysis, setting a new standard for player valuation in sports analytics.

The significantly lower RMSE values observed in this study compared to previous research highlight the superior accuracy of the models developed here as can be seen in table 18. The improvement in accuracy can be attributed to several key factors. First, the application of logarithmic transformation effectively normalized the distribution of player values which reduces the skewness often present in such datasets. This transformation ensures that the model is better at handling a wide range of values arguably leading to smaller prediction errors. Additionally, the models were less influenced by extreme values that could distort the predictions by identifying and removing outliers. This further lowered the RMSE.

Despite these improvements in RMSE, the R^2 values remained similar to those reported in other studies, with attackers being superior and the other positions being slightly lower. This is because R^2 measures the proportion of variance explained by the model, which can remain high even if there are large absolute errors in some cases. However, the lower RMSE in this study indicates that the model not only captures the overall variance but also achieves much more precise predictions, making it superior to the models in previous studies.

This study contributes to the academic field by advancing the integration of machine learning in football player valuation. Previous research often focusses on limited feature subsets or generic modeling approaches. This study demonstrates that a comprehensive position-based approach using the latest machine learning algorithms significantly enhances predictive accuracy. By incorporating features from all relevant categories and leveraging the latest machine learning methods such as PSO-SVR, CatBoost and XGBoost, the study addresses the research gap related to feature integration and model specificity. These findings extend the theoretical understanding of player valuation models by providing evidence of the advantages of ensemble and optimization-based approaches, setting a new standard for future research in football analytics.

The valuation models developed in this study offer significant advantages for football clubs. These models allow for enabling more informed and strategic decision-making across various aspects of club operations. The main advantages for football clubs using these models are discussed below.

One of the primary contributions of these models is their ability to support data-driven and objective recruitment decisions. By providing more precise player valuations, these models help clubs avoid overpaying for players whose market value may be inflated by media hype or market speculation because it focusses on objective data. The valuation models in this study provide a stable, consistent approach to valuing players which helps clubs avoid market-driven price inflation and make more calculated, data-backed investment decisions. By tracking player performance trends and market dynamics over time, the models enable clubs to anticipate shifts in player values and adjust their recruitment strategies accordingly, ensuring they remain competitive advantage.

Valuation models also enable clubs to forecast player values with greater precision. This significantly improves budgeting and financial planning. By using historical data and performance trends, these models can predict how a player's value will evolve over time. This allows clubs to plan their transfer and salary budgets accordingly. This foresight is crucial for clubs looking to balance short-term success with long-term financial sustainability, as it allows them to allocate resources more effectively.

Another key advantage of these models is their ability to pair player value with specific variables relevant to each position. Clubs can evaluate defenders based on metrics like tackles, interceptions and aerial duels, while midfielders might be assessed for passing accuracy and ball progression with more precision. By understanding which variables matter most for each position, clubs can refine their scouting efforts to focus on players who meet the specific tactical and performance needs of the team. This leads to more informed recruitment decisions and a stronger, more balanced squad.

Valuation models are particularly useful for identifying emerging talent, as they rely on objective performance data rather than subjective opinions or media-driven reputations. By analyzing a player’s metrics across leagues and competitions, clubs can spot young or lesser-known players with high potential before their market value rises. This proactive approach gives clubs a competitive edge in the transfer market, allowing them to acquire promising talents early and develop them into key contributors.

The last major advantage is that these models offer clubs valuable insights into when to sell or retain players. This enables clubs for more strategic transfer decisions. By predicting when a player's value is likely to peak or decline based on performance trends, age and other relevant features identified in this study, clubs can time their sales to maximize transfer income. Similarly, they can identify when to keep a player whose value is expected to rise. This ensures they maintain a competitive squad without unnecessarily offloading key assets (players). This strategic approach to transfer decision-making helps clubs optimize both sporting success and financial returns.

Nevertheless, it is also important to acknowledge the limitations of this study. The relatively small size of the dataset may have constrained the performance of the meta-model, as larger datasets typically allow for more robust training and validation of ensemble models. Additionally, the logarithmic transformation applied to the data may have influenced the interpretation of player values even though it was necessary due to the distribution. These limitations suggest

that further research with larger and more diverse datasets can be helpful to fully validate the models' effectiveness and possibly further enhance the accuracy. For instance, with player fitness data which was not considered in this research. Football clubs nowadays have direct access to this data from every player of their club.

In conclusion, this study demonstrates that the best-performing position-based player valuation models built with the latest machine learning methods do indeed outperform existing models. The PSO-SVR model in particular stands out for its high accuracy and robustness across multiple player positions. These findings also confirm that a position-specific approach in combination with the latest machine learning techniques provides a more precise and reliable method for predicting football player values. This research not only contributes to the academic understanding of the latest algorithms used for prediction models. It also offers practical implications for football clubs looking to enhance their decision-making processes.

Appendices

Appendix 1: Overview of previous created player valuation models

Table 1: Overview of related works with their research purpose, data sources, features used and methods

Reference	Research objective	Data sources	Features	Methods
He et al. (2015)	Estimation of players performance and market value, and relationship between player's performance and market value by regression model	Transfer market, WhoScored, Eurpoean Football Database, Garter	Transfer fee, performance assessments, age, contract duration	Lasso regression
Majewski et al. (2016)	Estimation of player's market value and identifying the determing factors of market value by regression model	Transfer market	5 Human capital factors (e.g., age), 5 Productivity factors (e.g., goals scored), 4 Organizational capital factors (e.g., total time)	Linear regression
Müller et al. (2017)	Estimation of player's market value by regression model	Google, Reddit, Transfer market, WhoScored, Wikipedia, YouTube	1 Player valuation (e.g., market value), 3 Player characteristics (e.g., age), 16 Player performance (e.g., minutes played), 4 Player popularity (e.g., Wikipedia page view)	Linear regression
Behravan and Razavi (2020)	Estimation of player's market value by regression model	SOFIFA	55 attributes: (Physical, Attacking, Movement, Skill, Defensive, Mentality, Power, General), 5 features for goalkeepers, 32 features for attackers, 30 features for defenders, 28 features for midfielders chosen by PSO clustering	Particle Swarm Optimization (PSO) SVR, Gery Wolf Optimizer (GWO) SVR, Inclined Planes System Optimization (IPO) SVR, Whale Optimization Algorithm (WOA) SVR
Yigit et al. (2020)	Estimation of player's market value by regression model	Football Manager, Transfer market	49 attributes: technical, mental, physical and goalkeeping	Linear regression, ridge regression, lasso regression, principal component regression, random forest, XGBoost
Lee et al. (2022)	Prediction of player's market value using Bayesian Ensemble Approach	SOFIFA, WhoScored	Top 20 attributes from feature selection	Reguralized Linear regression, Gradient Boosting decision tree, LightGBM, Hyperparameter optimization

Appendix 2: Overview of features in dataset

Table 2: Description of features in dataset: monetary value, player characteristics, player performance, player potential, team features and crowd-judgement

Types of features	Features	Description of features	Data type	Range of possible values
Monetary value	Value Transfermarkt	Market value of a player according to Transfermarkt.com	Ratio	No limitation
	Value SOFIFA	Market value of a player according to SOFIFA	Ratio	No limitation
	Combined value	Aggregated value of a player	Ratio	No limitation
	Wage	Weekly salary of a player from affiliated club	Ratio	No limitation
	Release clause	Buyout clause of player to transfer	Ratio	No limitation
Player characteristics	Nationality	Nationality of a player	Nominal	0 or 1
	International reputation	Reputation of a player's country of origin	Ordinal	1, 2, 3, 4 or 5
	Club name	Name of the player's club	Nominal	0 or 1
	Club joined	When the player joined the club	Ratio	No limitation
	Contract valid until year	Year until the player's contract is valid	Ratio	No limitation
	Overall rating	Overall rating based on ability features	Interval	1-99
	Age	Age of a player	Ratio	No limitation
	Date of birth	Date of birth of a player	Scale	No limitation
	Height	Height of a player	Ratio	No limitation
	Weight	Weight of a player	Ratio	No limitation
	Positions	Positions of player	Nominal	0 or 1
	Rating per position	Overall rating per position	Interval	1-99
	Best positions	Best positions	Nominal	0 or 1
	Position category	Category of which the player's positions belong to (attackers, midfielders, defenders, goalkeepers)	Nominal	0 or 1
	Preferred foot	Preferred foot of a player (right or left)	Nominal	0 or 1
	Weak foot	Weak foot of a player (right or left)	Ordinal	1, 2, 3, 4 or 5
	Skill moves	Ability to perform skill moves	Ordinal	1, 2, 3, 4 or 5
	Attacking work rate	Attacking work rate of a player	Ordinal	0, 1 or 2
	Defensive work rate	Defensive work rate of a player	Ordinal	0, 1 or 2
	Ability features	Aggegrated sum of 35 ability features from SOFIFA	Interval	1-99
Player performance	Appearances	Appearances of a player in eredivisie season 2023/2024	Interval	0-34
	Minutes played	Minutes played in eredivisie season 2023/2024	Ratio	No limitation
	Goals	Goals of a player in eredivisie season 2023/2024	Ratio	No limitation
	Assists	Assists of a player in eredivisie season 2023/2024	Ratio	No limitation
	Yellow cards	Yellow cards of a player in eredivisie season 2023/2024	Interval	1-877
	Red cards	Red cards of a player in eredivisie season 2023/2024	Interval	1-34
	Shots per game	Shots per game	Ratio	No limitation
	Passing accuracy	Passing succes percentage of a player in eredivisie season 2023/2024	Interval	0-100
	Aerials won	Aerial duels won per game	Ratio	No limitation
	Key passes	Key passes per game	Ratio	No limitation
	Dribbles	Dribbles per game	Ratio	No limitation
	Fouled	Fouled per game	Ratio	No limitation
	Disposessed	Ball dispossessions per game	Ratio	No limitation
	Tackles	Tackles per game	Ratio	No limitation
	Interceptions	Interceptions per game	Ratio	No limitation
	Fouls	Fouls per game	Ratio	No limitation
	Clearances	Clearances per game	Ratio	No limitation
	Blocks	Blocks per game	Ratio	No limitation
	Own goals	Own goals of a player in eredivisie season 2023/2024	Ratio	No limitation
	Average passes	Average passes per game	Ratio	No limitation
	Man of the match	Man of the match award of a player in eredivisie season 2023/2024	Ratio	No limitation
Player potential	Potential	Potential rating from SOFIFA based on ability features, age and international reputation	Interval	1-99
Team features	Goal acquisition	Total number of goal scored by team in season	Ratio	No limitation
	Goal against	Total goals scored by the opposite team in season	Ratio	No limitation
	Goal difference	Goal acquisition-Goal against	Ratio	No limitation
	Victory point	Total victory point in season	Ratio	No limitation
	Win point	The number of wins of the season	Interval	1-34
	Draw point	The number of draws of the season	Interval	1-34
	Lose point	The number of losses of the season	Interval	1-34
	Team standing	Team ranking at the end of the season	Ordinal	1-18
Crowd-judgement	Rating	Avarage rating from crowd-judgement and experts at the end of the season	Interval	0-10

Appendix 3: Overview of ability features from SOFIFA

Table 3: Description, calculation and range of ability features from SOFIFA

Calculated ability features	Formula	Range of possible values
PAC	(Sprint Speed + Acceleration)/2	1–99
SHO	(Finishing + Long Shots + Shot Power)/3	1–99
PAS	(Crossing + Short Passing + Long Passing)/3	1–99
DRI	(Ball Control + Agility + Balance)/3	1–99
DEF	(Marking + Tackling + Strength)/3	1–99
PHY	(Strength + Stamina + Jumping)/3	1–99
Attacking	Crossing + Finishing + Heading Accuracy + Short Passing + Volleys	5–495
Skill	Dribbling + Curve + FK Accuracy + Long Passing + Ball Control	5–495
Movement	Acceleartion + Agility + Sprint Speed + Reactions+ Balance	5–495
Power	Shot Power + Jumping + Stamina + Strength + Long Shots	5–495
Defending	Marking + Sliding Tackle + Standing Tackle	3–297
Mentality	Aggression + Reactions + Positioning + Interceptions + Vision + Composure	6–594
Goalkeeping	GK Positioning + GK Diving + GK Handling + GK Kicking + GK Reflexes	5–495
Overall Rating	Overall rating in position	1–99
BOV	Overall rating in best position	1–99
Base stats	PAC+SHO+PAS+DRI+DEF+PHY	6–594
Total stats	Sum of total 35 ability elements	39–3500

Appendix 4: Description of player position types

Table 4: Description of player position types

Position category	Position sub-category	Position abbreviation	Position description
Attackers	Striker	LS ST RS	Left striker Striker Right striker
	Forward	LF CF RF	Left forward Center forward Right forward
	Winger	LW RW	Left winger Right winger
Midfielders	Wide midfielder	LM RM	Left midfielder Right midfielder
	Attacking midfielder	LAM CAM RAM	Left attacking midfielder Central attacking midfielder Right attacking midfielder
	Central midfielder	LCM CM RCM	Left central midfielder Central midfielder Right central midfielder
	Defensive midfielder	LDM CDM RDM	Left defensive midfielder Central defensive midfielder Right defensive midfielder
Defenders	Wingback	LWB RWB	Left wing back Right wing back
	Full back	LB RB	Left back Right back
	Center back	LCB CB RCB	Left center back Center back Right center back
Goalkeepers	Goalkeeper	GK	Goalkeeper

Appendix 5: Selected features from Pearson's correlation analysis for linear regression models

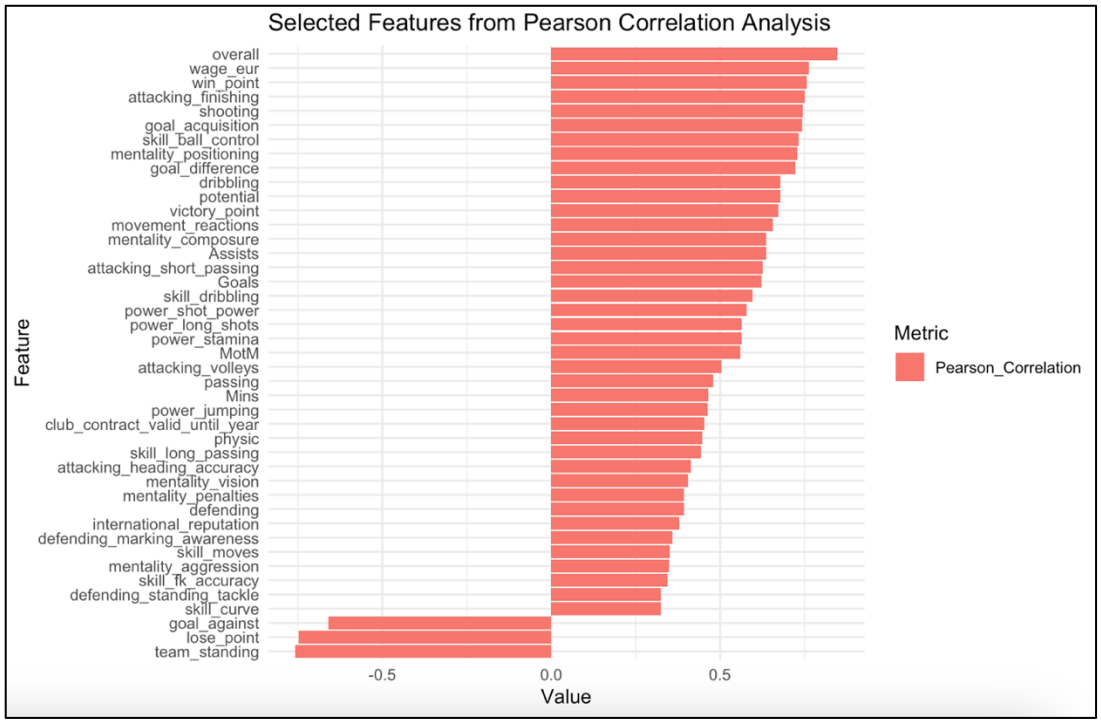

Figure 5: Selected features from Pearson's correlation analysis for linear regression models for attackers

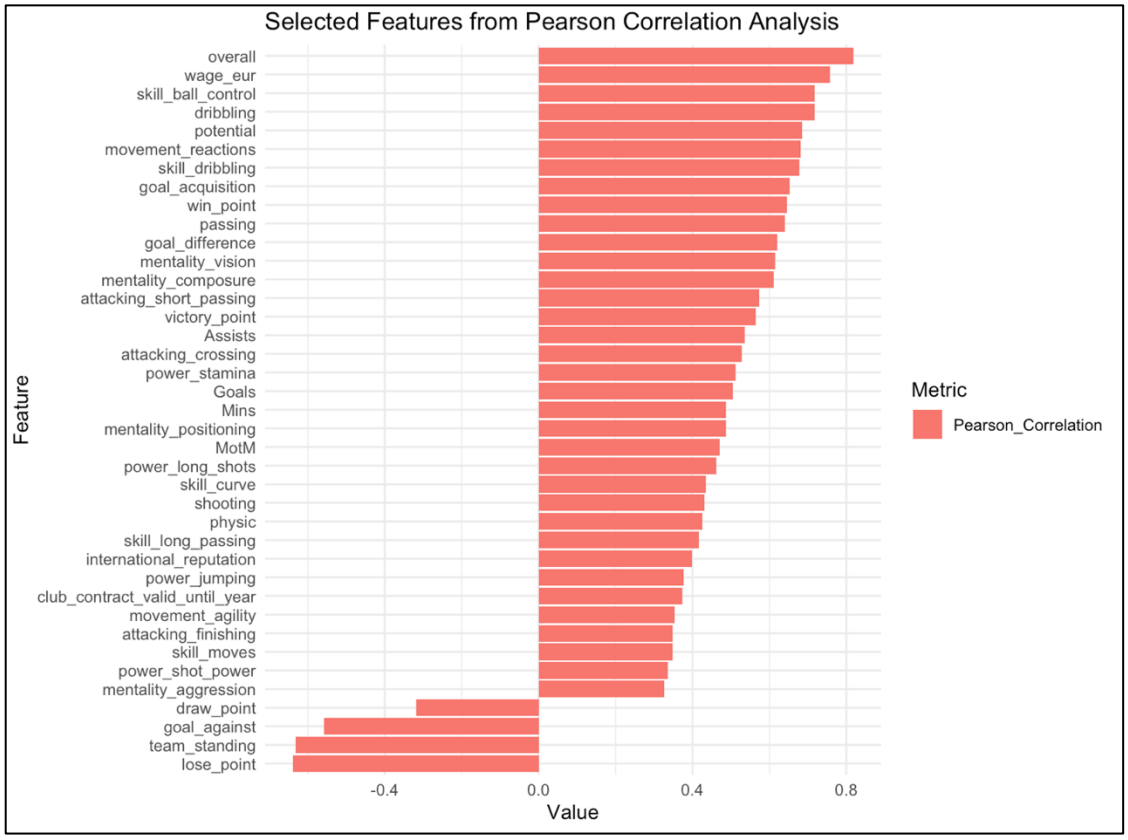

Figure 6: Selected features from Pearson's correlation analysis for linear regression models for midfielders

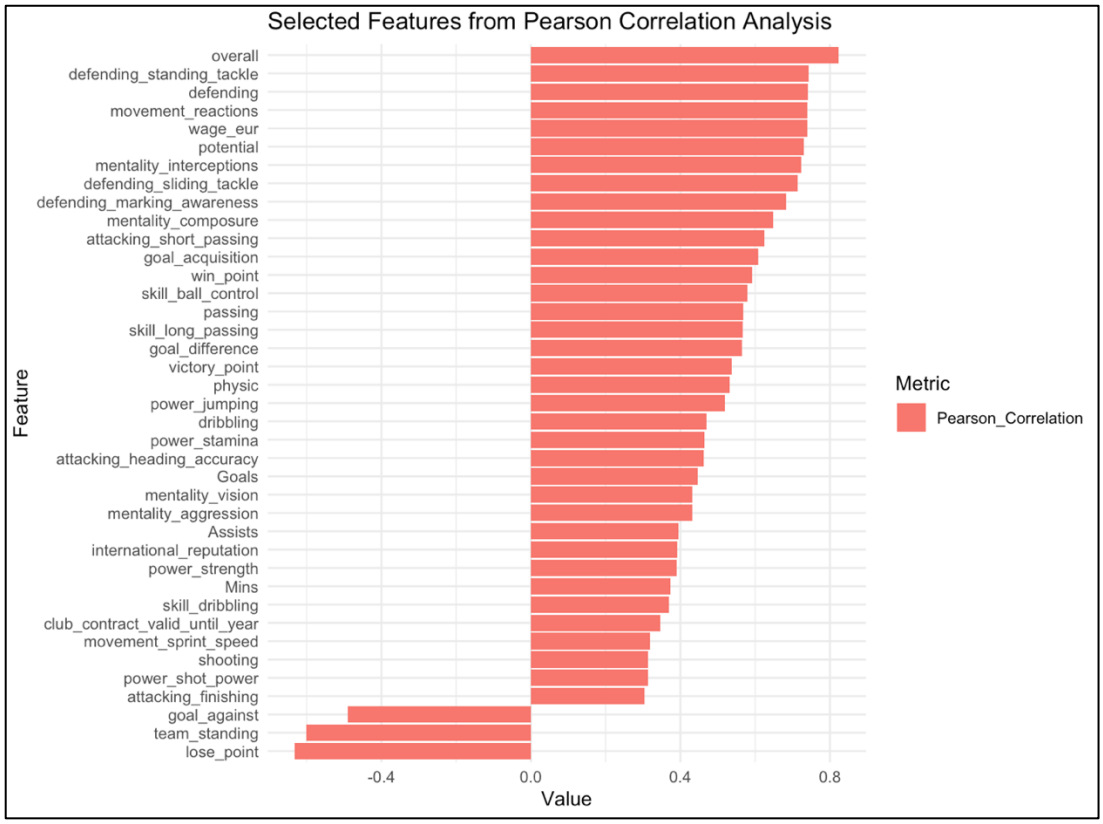

Figure 7: Selected features from Pearson's correlation analysis for linear regression models for defenders

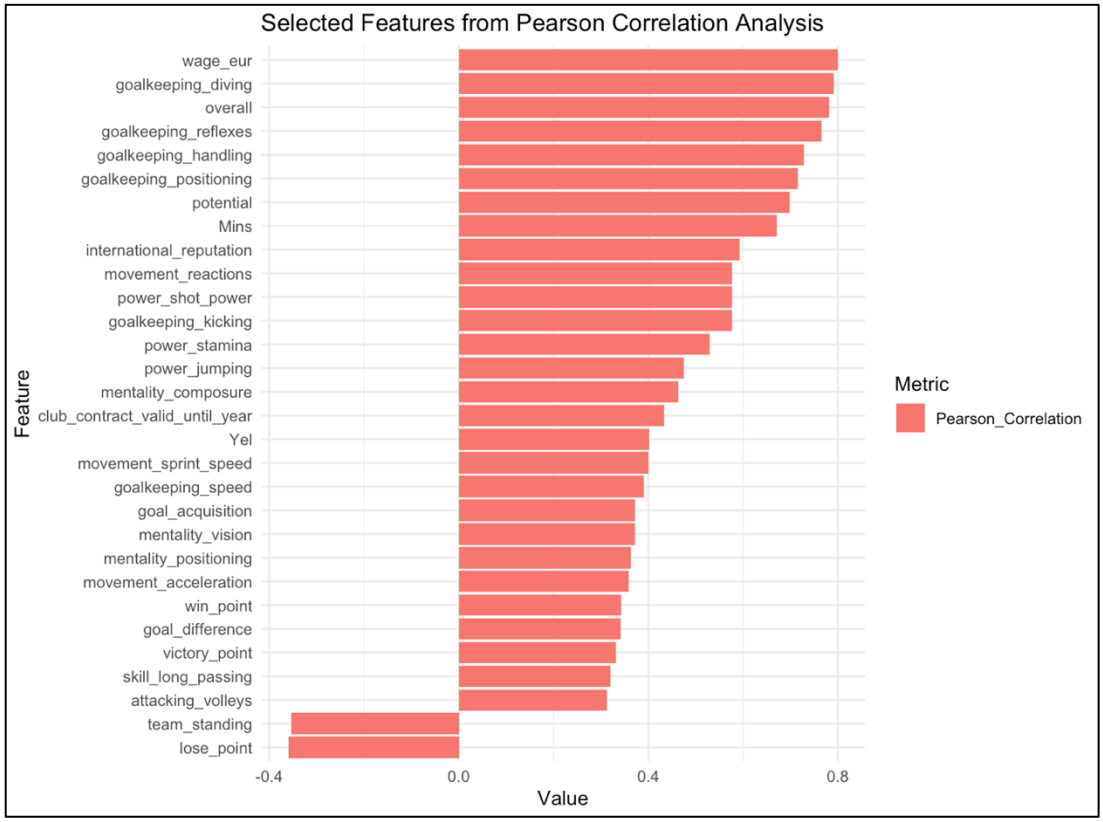

Figure 8: Selected features from Pearson's correlation analysis for linear regression models for goalkeepers

Appendix 6: Selected features from mean SHAP values analysis for linear regression models

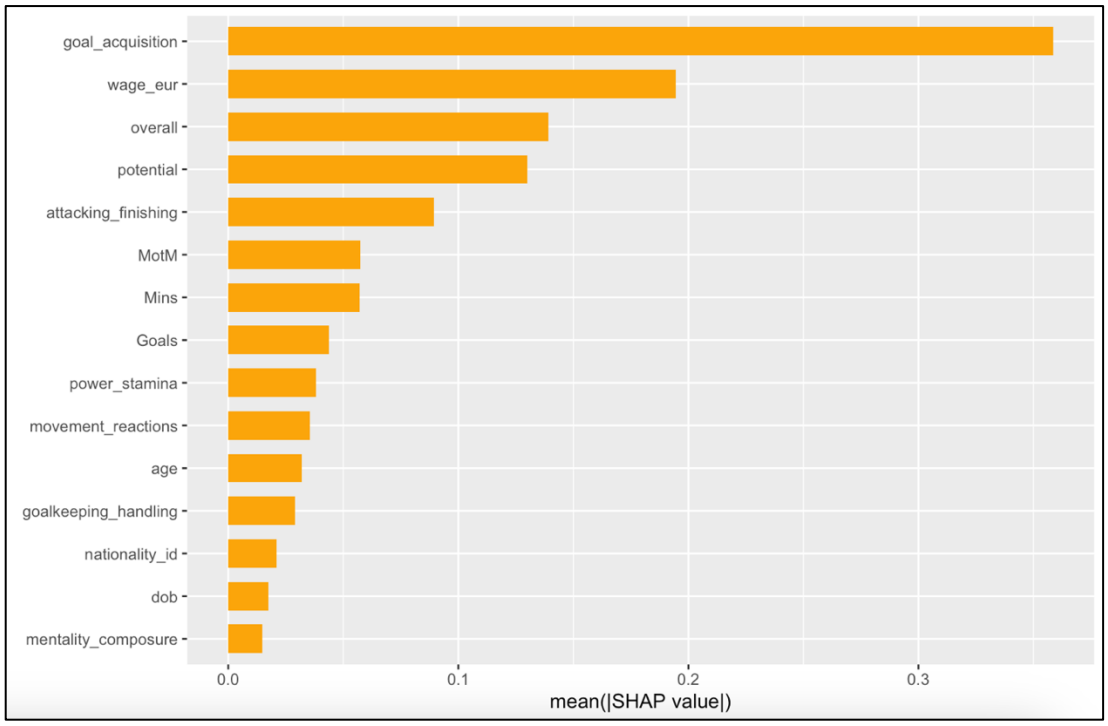

Figure 9: Selected features mean SHAP values analysis for linear regression models for attackers

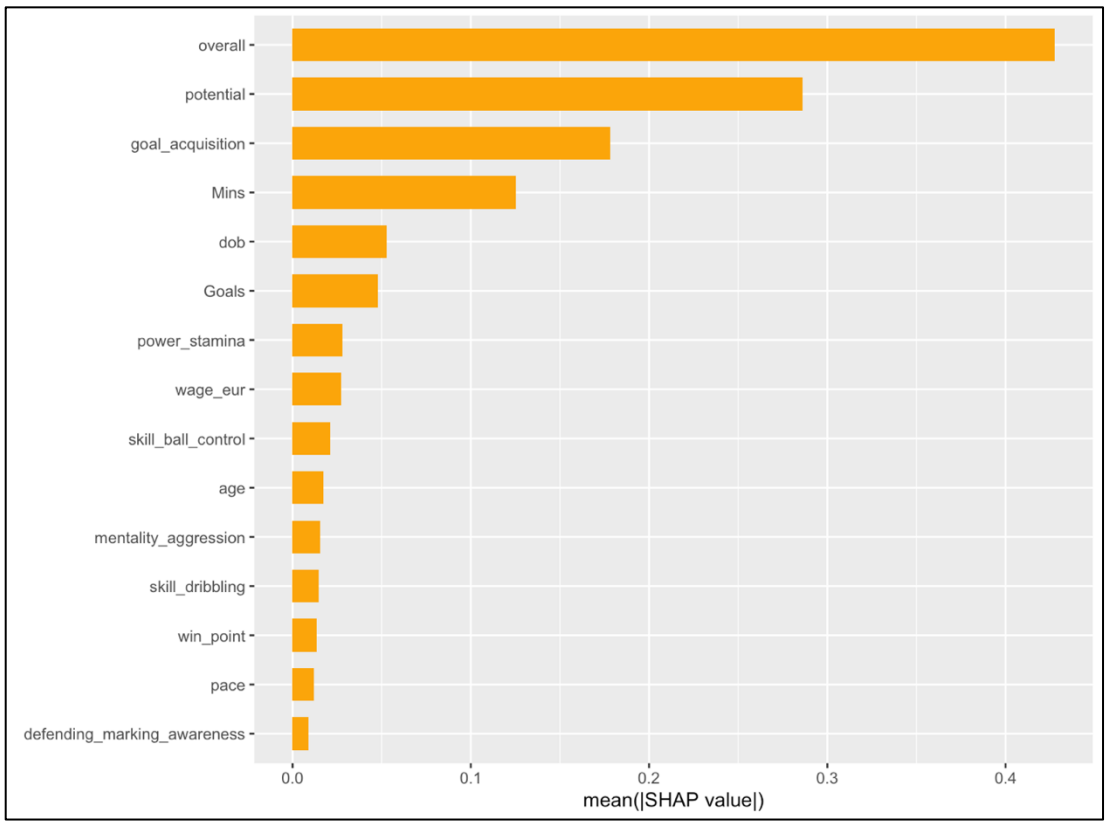

Figure 10: Selected features mean SHAP values analysis for linear regression models for midfielders

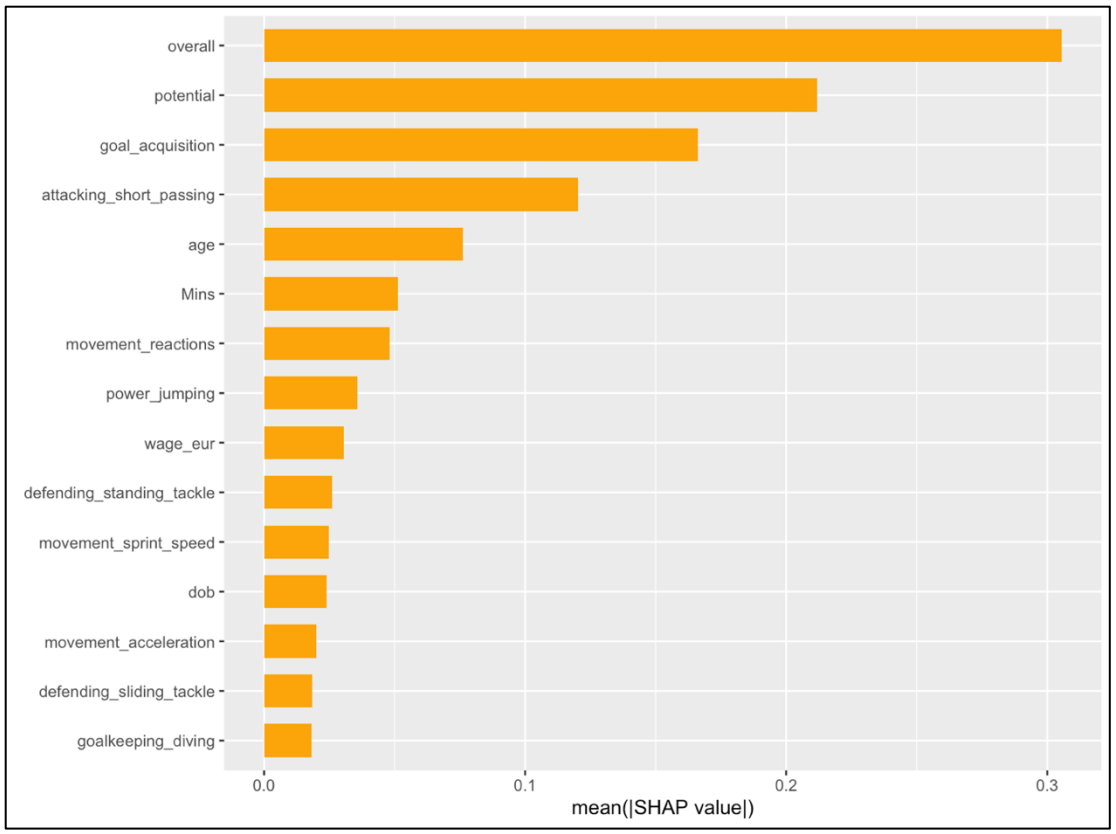

Figure 11: Selected features mean SHAP values analysis for linear regression models for defenders

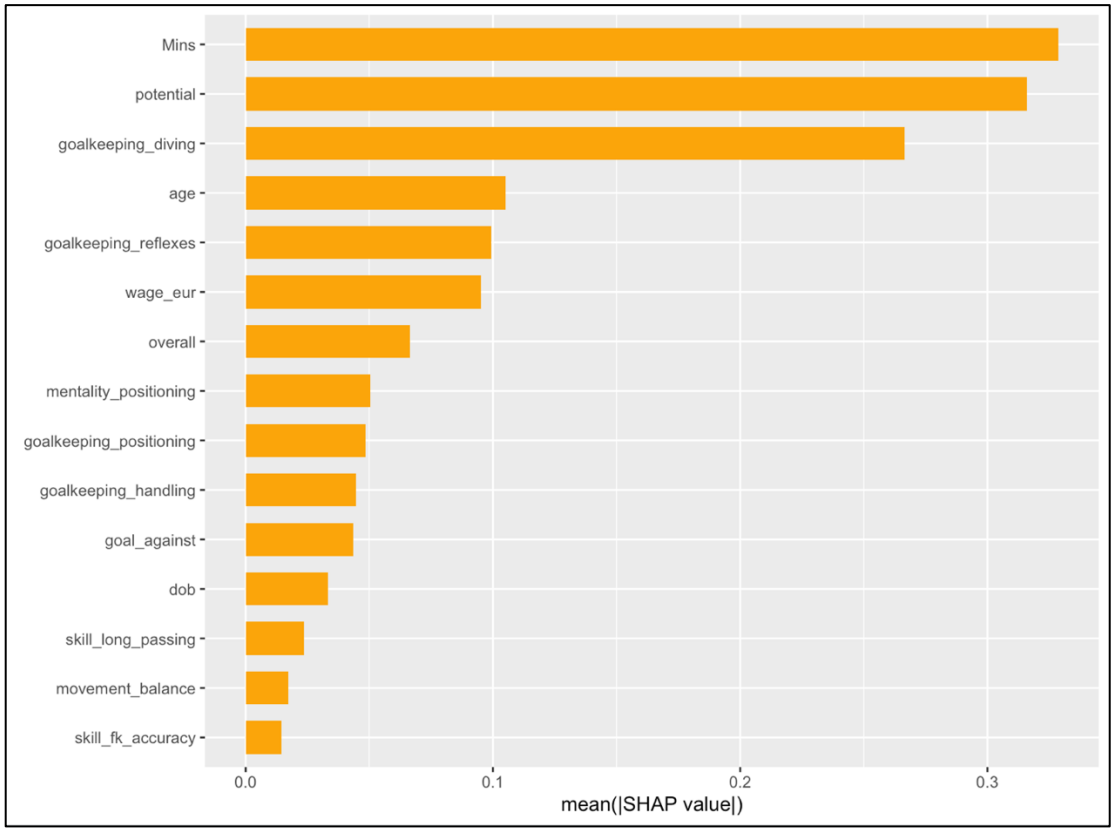

Figure 12: Selected features mean SHAP values analysis for linear regression models for goalkeepers

Appendix 7: SHAP summary plot with contribution of each feature to the prediction for PSO with SVR model

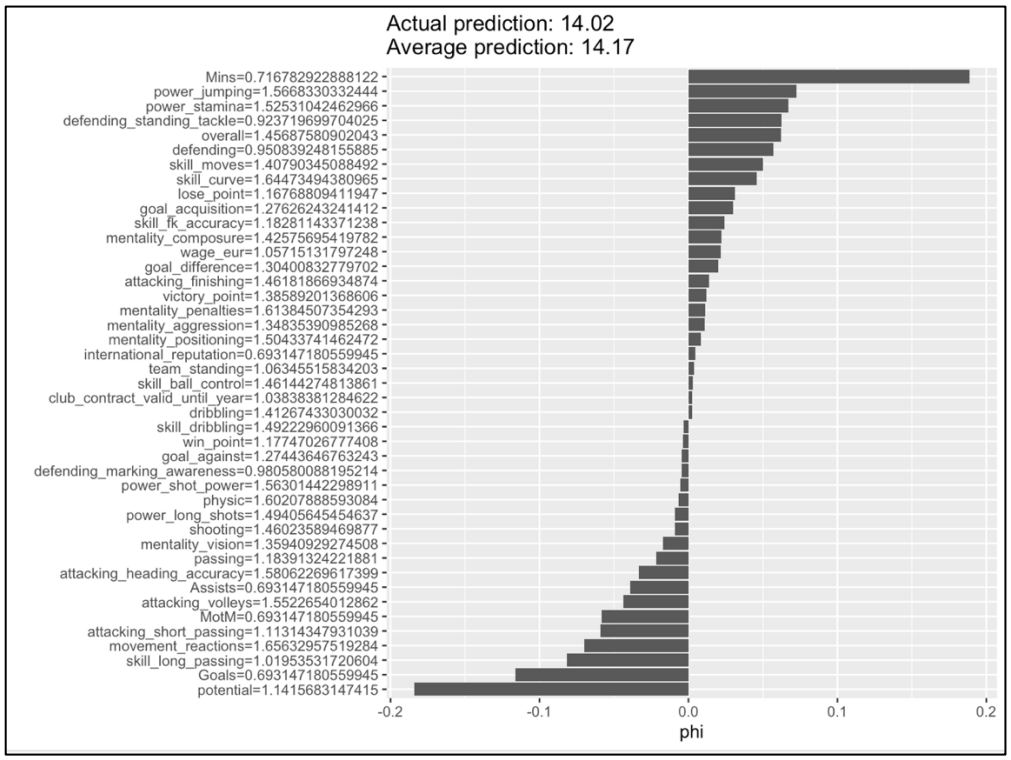

Figure 13: SHAP summary plot with contribution of each feature to the prediction for PSO with SVR model for attackers

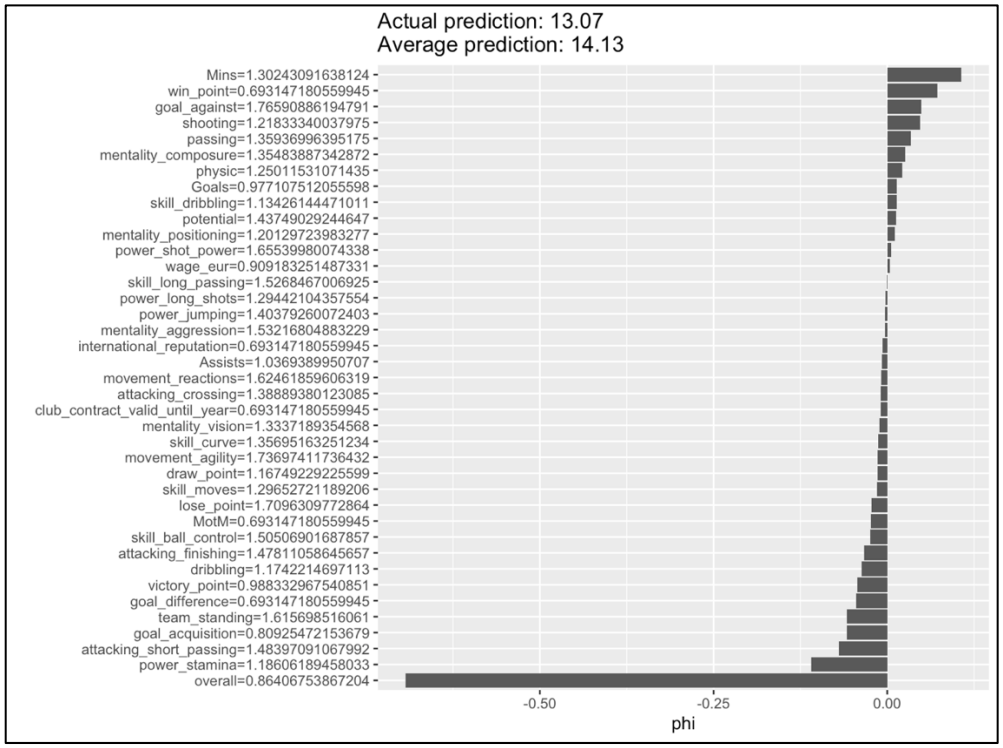

Figure 14: SHAP summary plot with contribution of each feature to the prediction for PSO with SVR model for midfielders

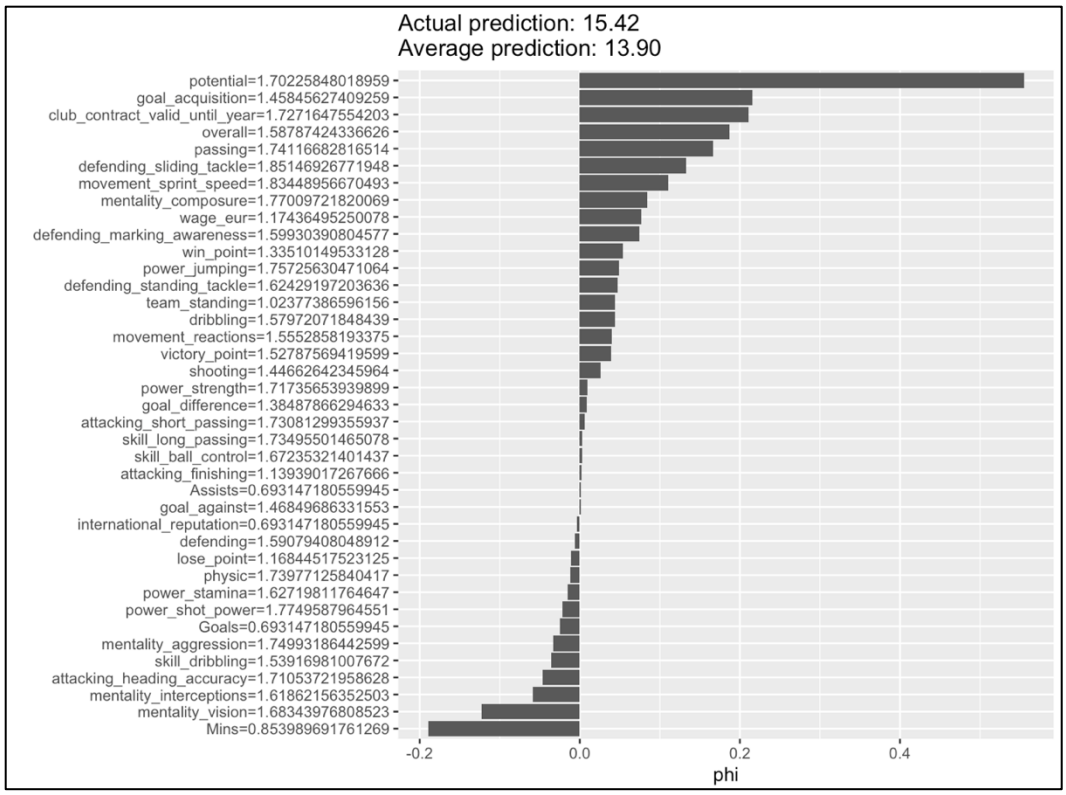

Figure 15: SHAP summary plot with contribution of each feature to the prediction for PSO with SVR model for defenders

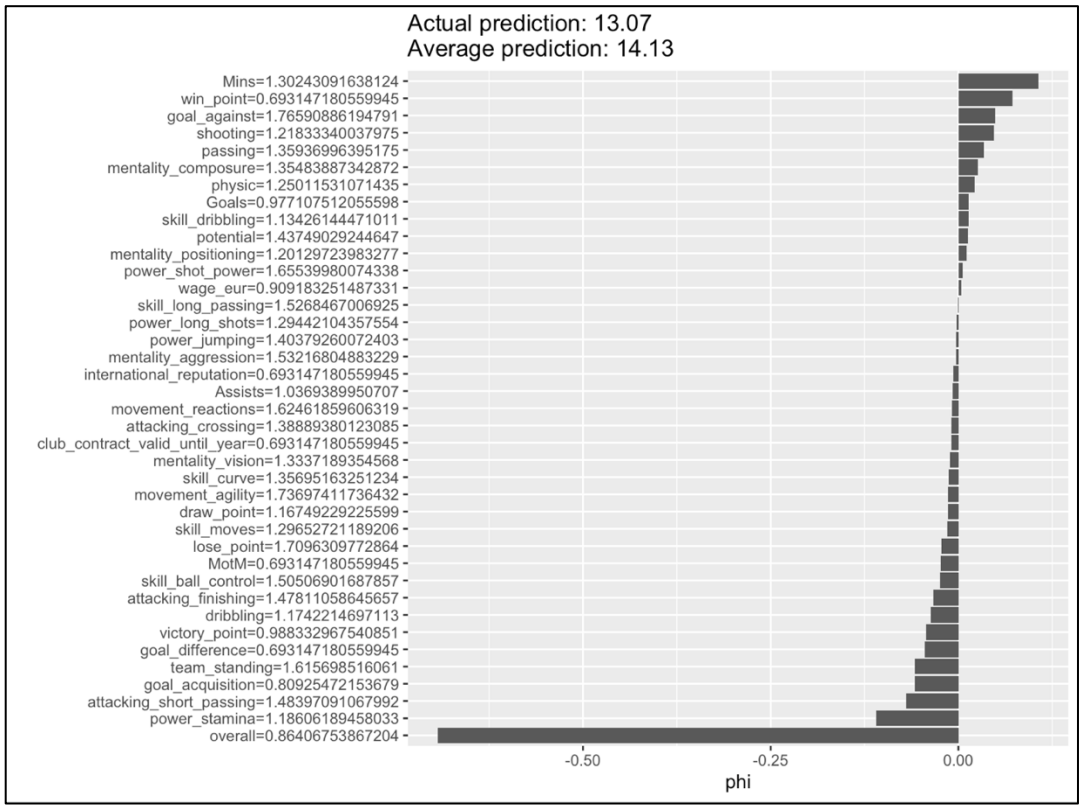

Figure 16: SHAP summary plot with contribution of each feature to the prediction for PSO with SVR model for goalkeepers

Appendix 8: Top 10 feature selection and importance plot for LightGBM model with Bayesian Optimization

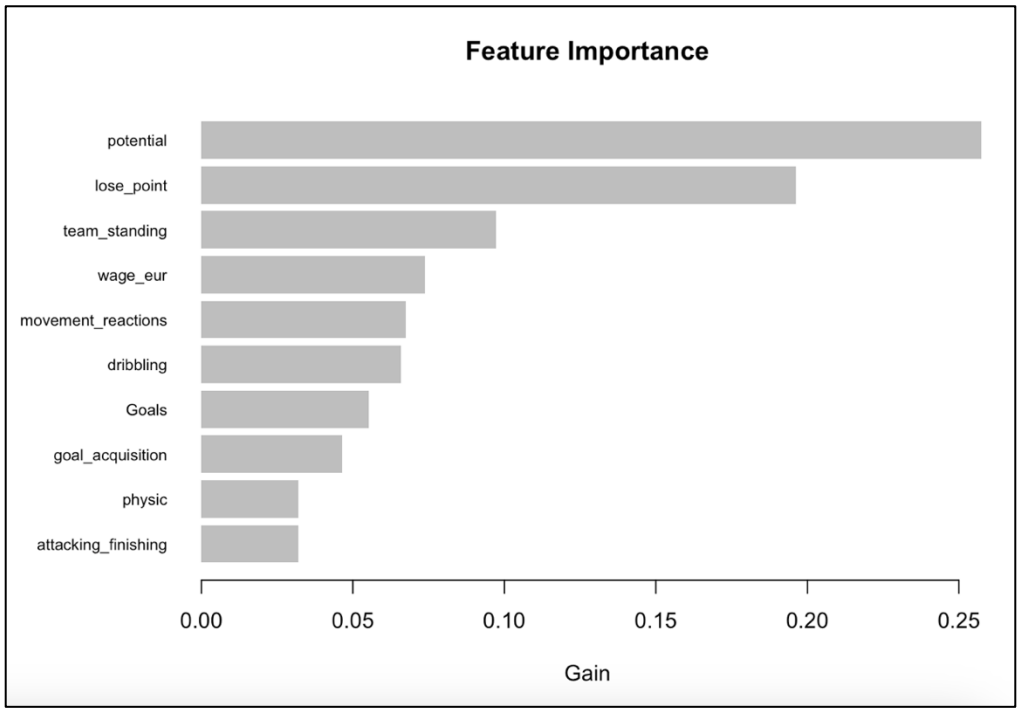

Figure 17: Top 10 feature selection and importance plot for LightGBM model with Bayesian Optimization for attackers

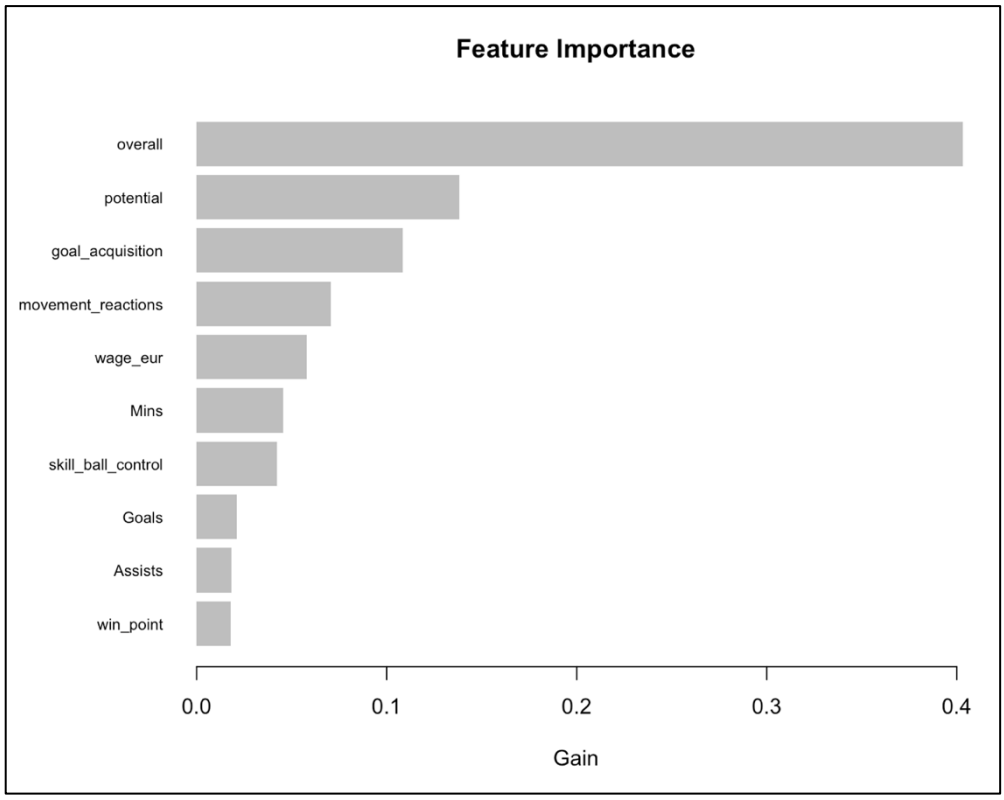

Figure 18: Top 10 feature selection and importance plot for LightGBM model with Bayesian Optimization for midfielders

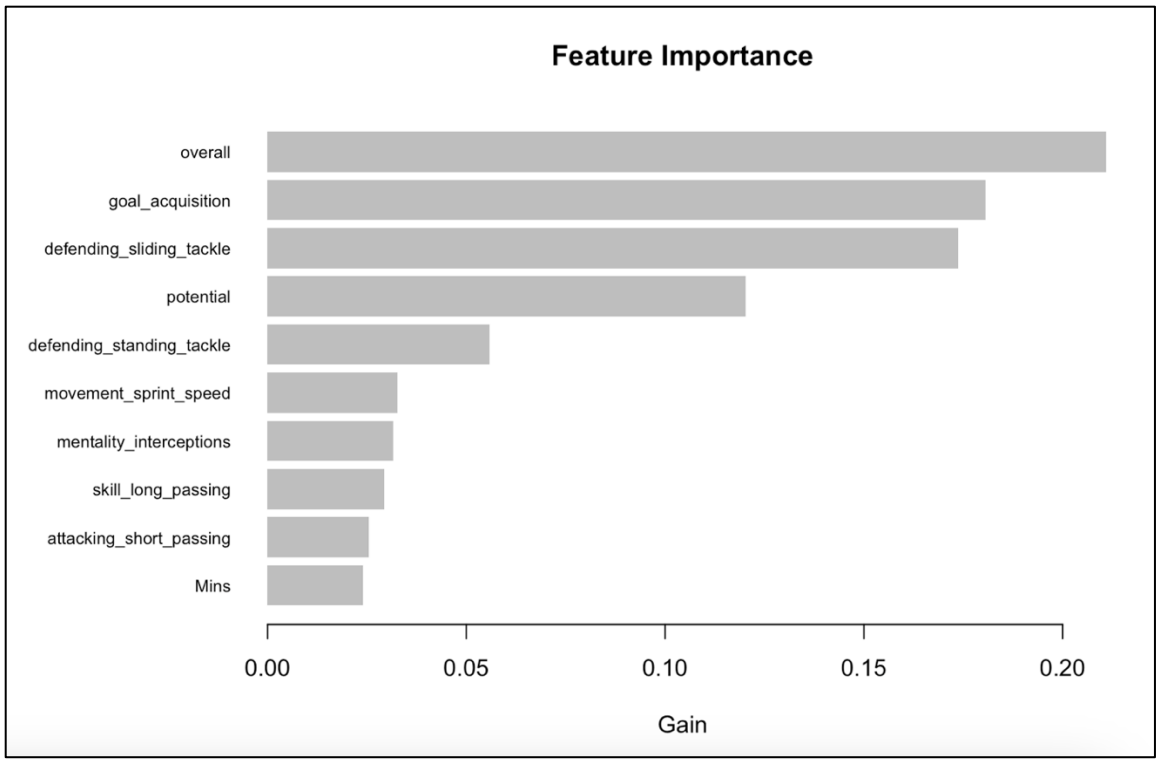

Figure 19: Top 10 feature selection and importance plot for LightGBM model with Bayesian Optimization for defenders

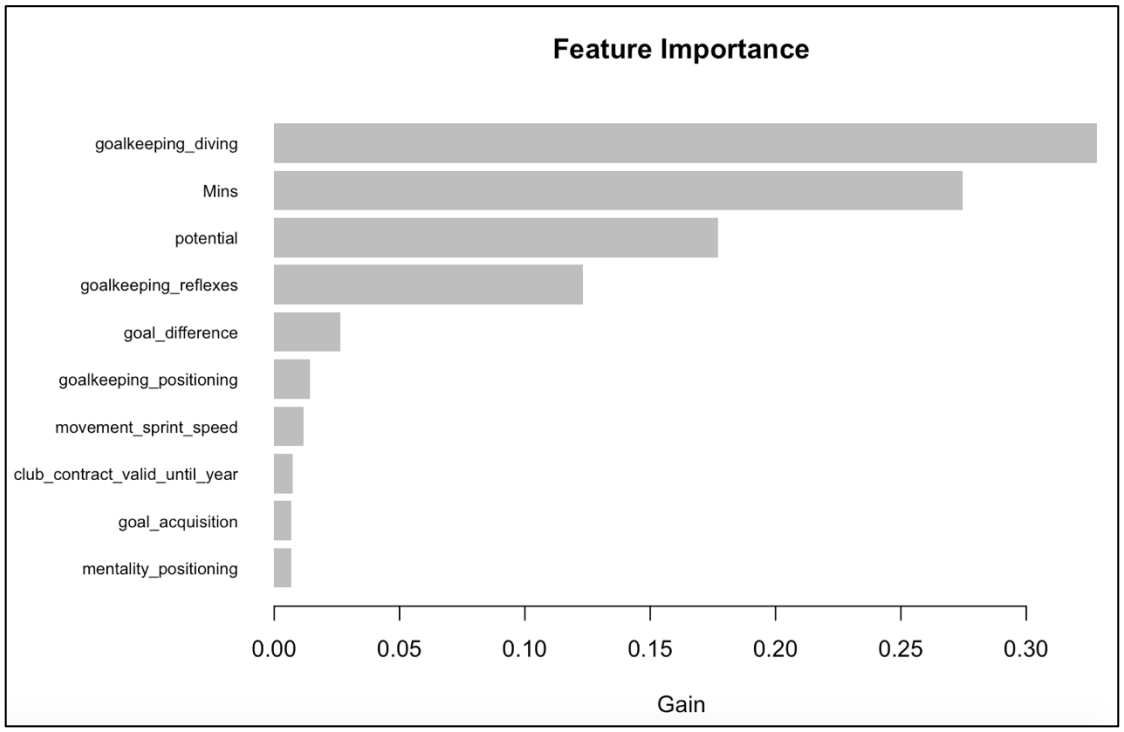

Figure 20: Top 10 feature selection and importance plot for LightGBM model with Bayesian Optimization for goalkeepers

Appendix 9: Top 10 feature selection and importance plot for XGBoost model with Bayesian Optimization

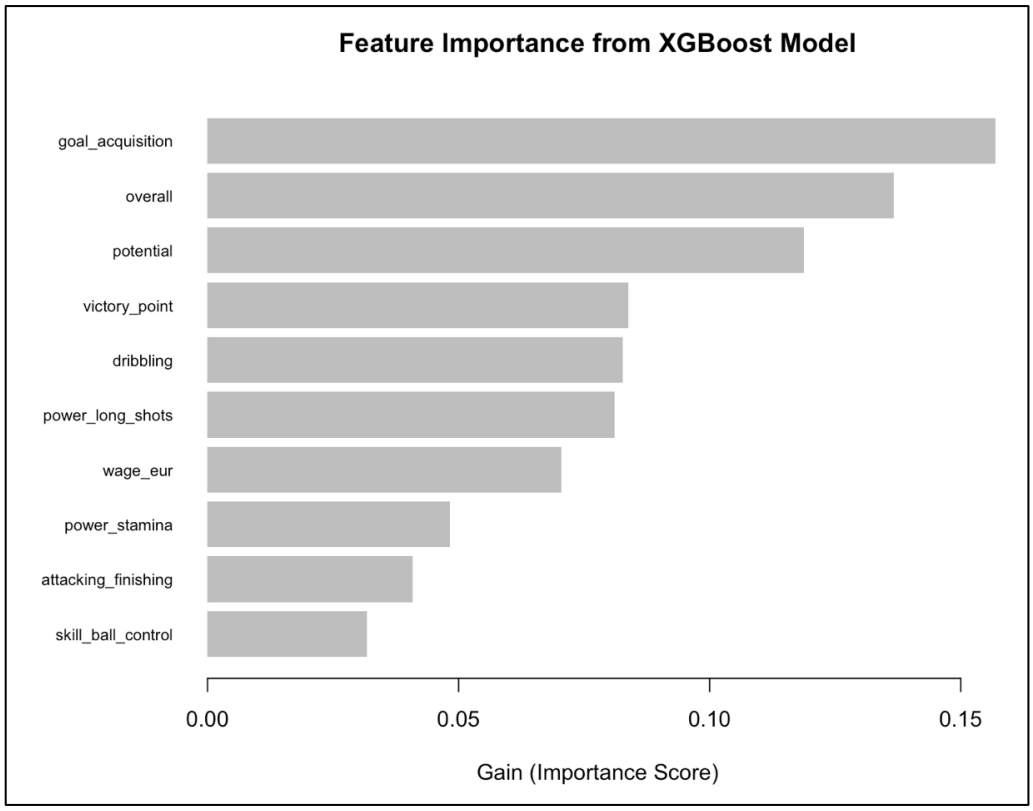

Figure 21: Top 10 feature selection and importance plot for XGBoost model with Bayesian Optimization for attackers

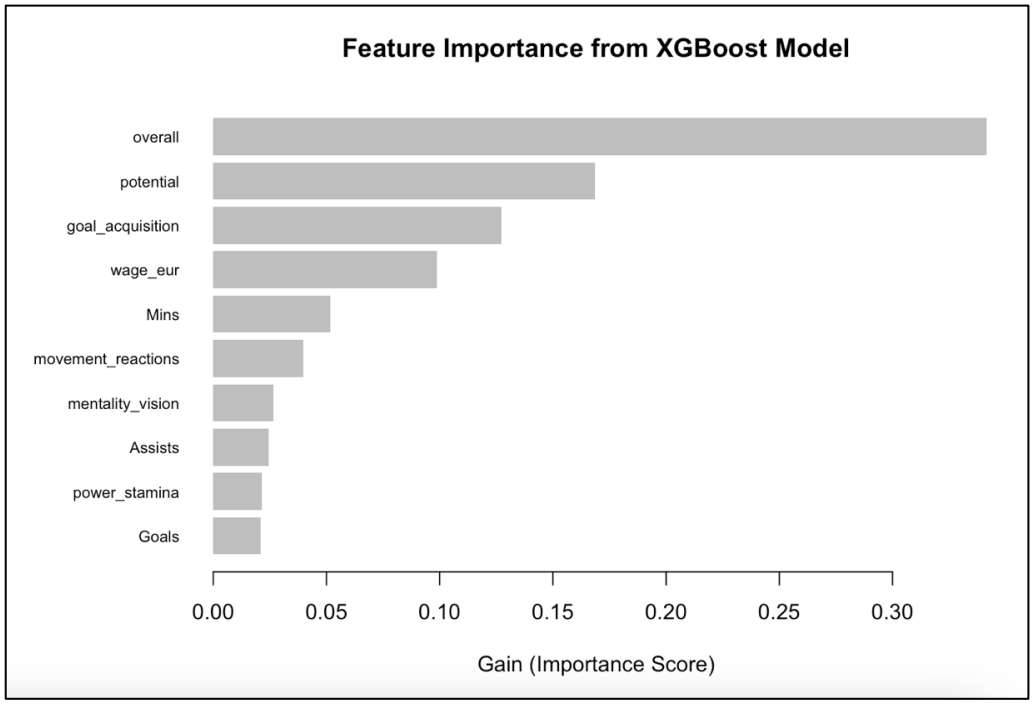

Figure 22: Top 10 feature selection and importance plot for XGBoost model with Bayesian Optimization for midfielders

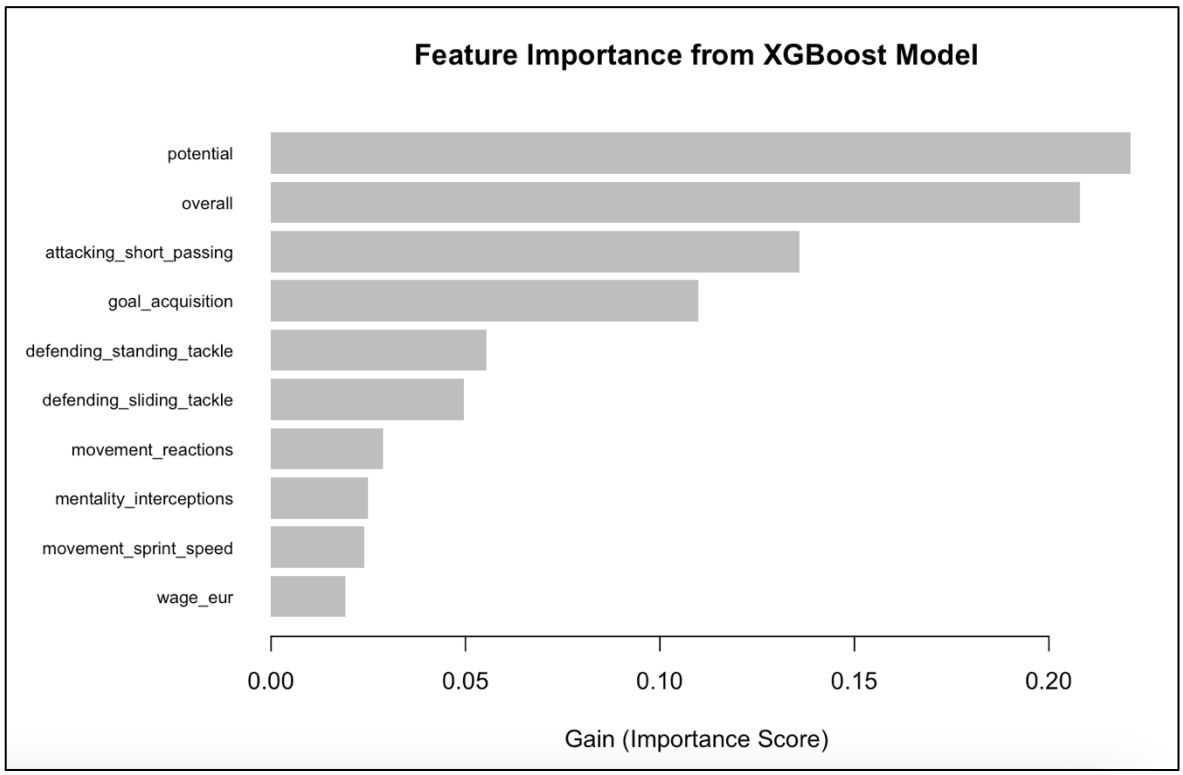

Figure 23: Top 10 feature selection and importance plot for XGBoost model with Bayesian Optimization for defenders

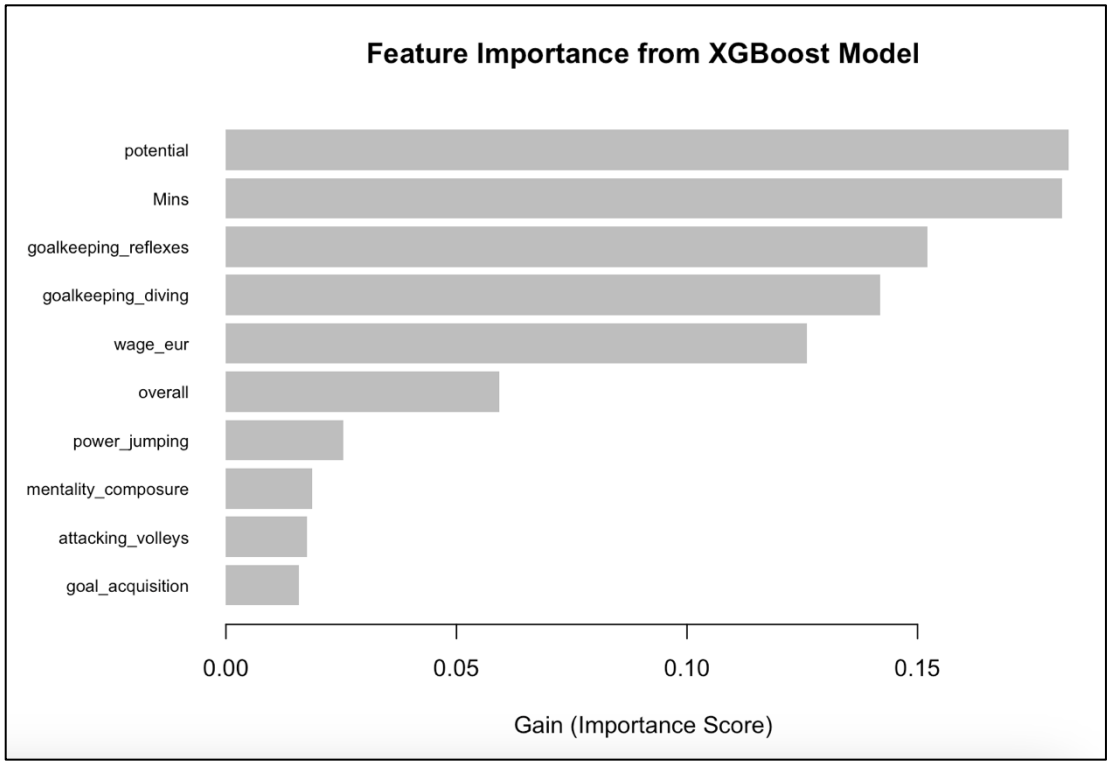

Figure 24: Top 10 feature selection and importance plot for XGBoost model with Bayesian Optimization for goalkeepers

Appendix 10: Feature selection and importance plot for CatBoost model with Bayesian Optimization

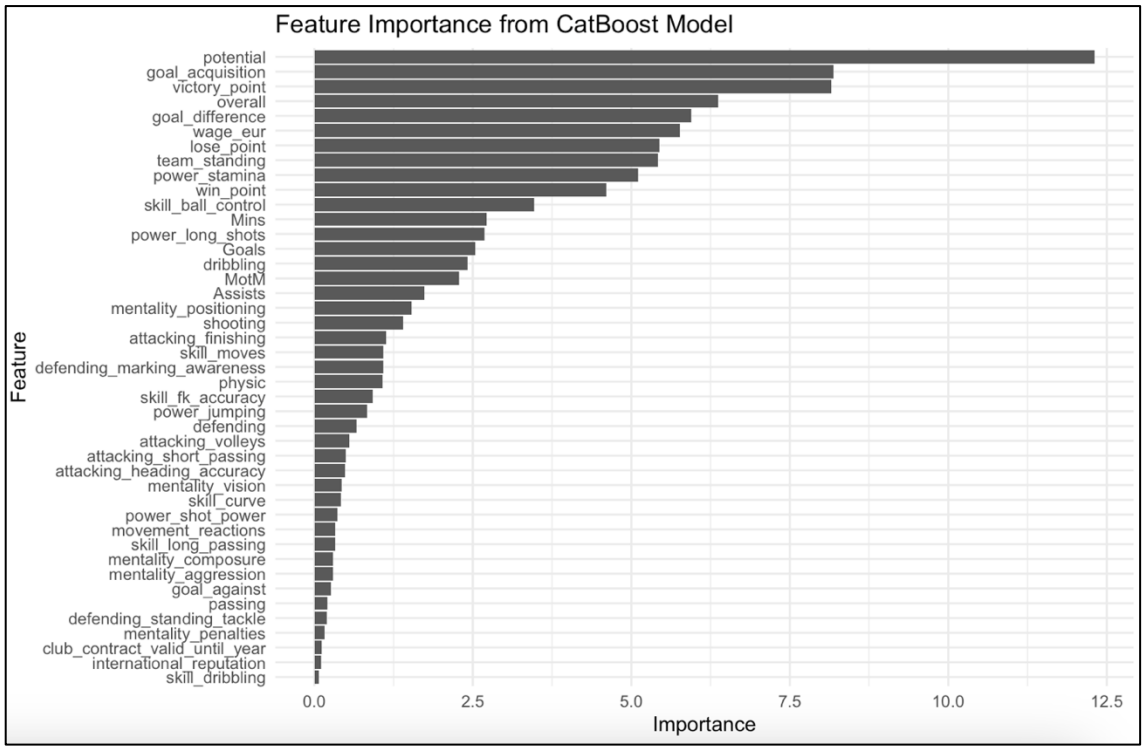

Figure 25: Feature selection and importance plot for CatBoost model with Bayesian Optimization for attackers

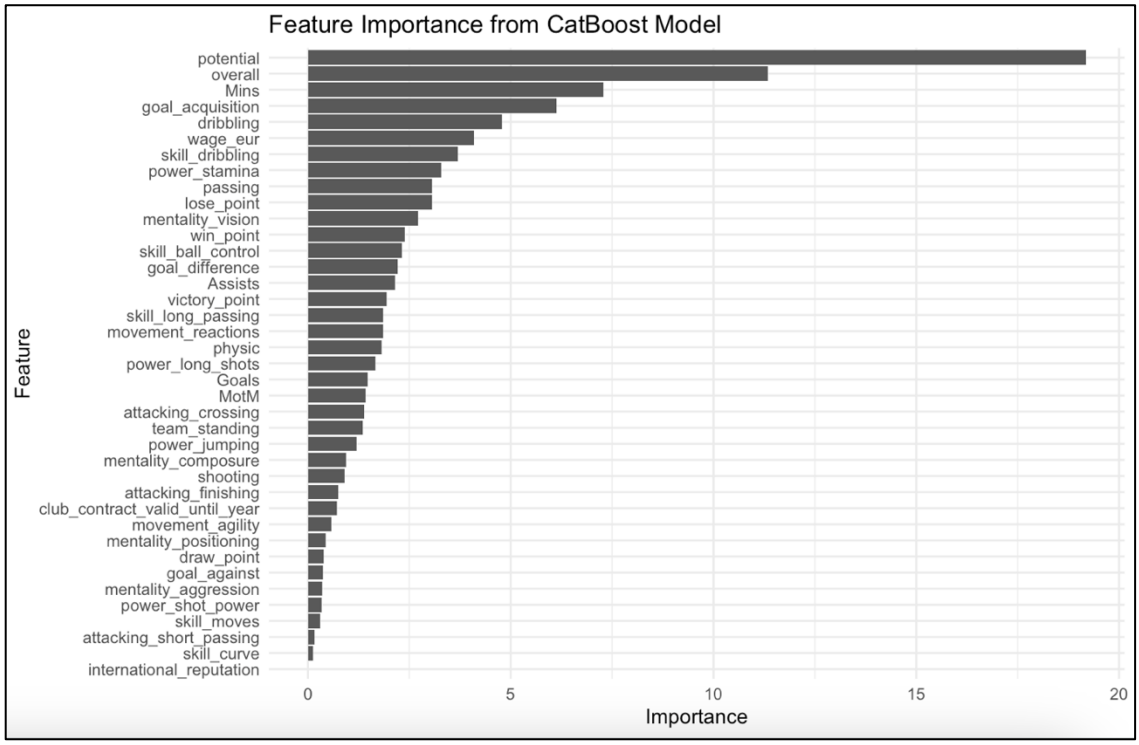

Figure 26: Feature selection and importance plot for CatBoost model with Bayesian Optimization for midfielders

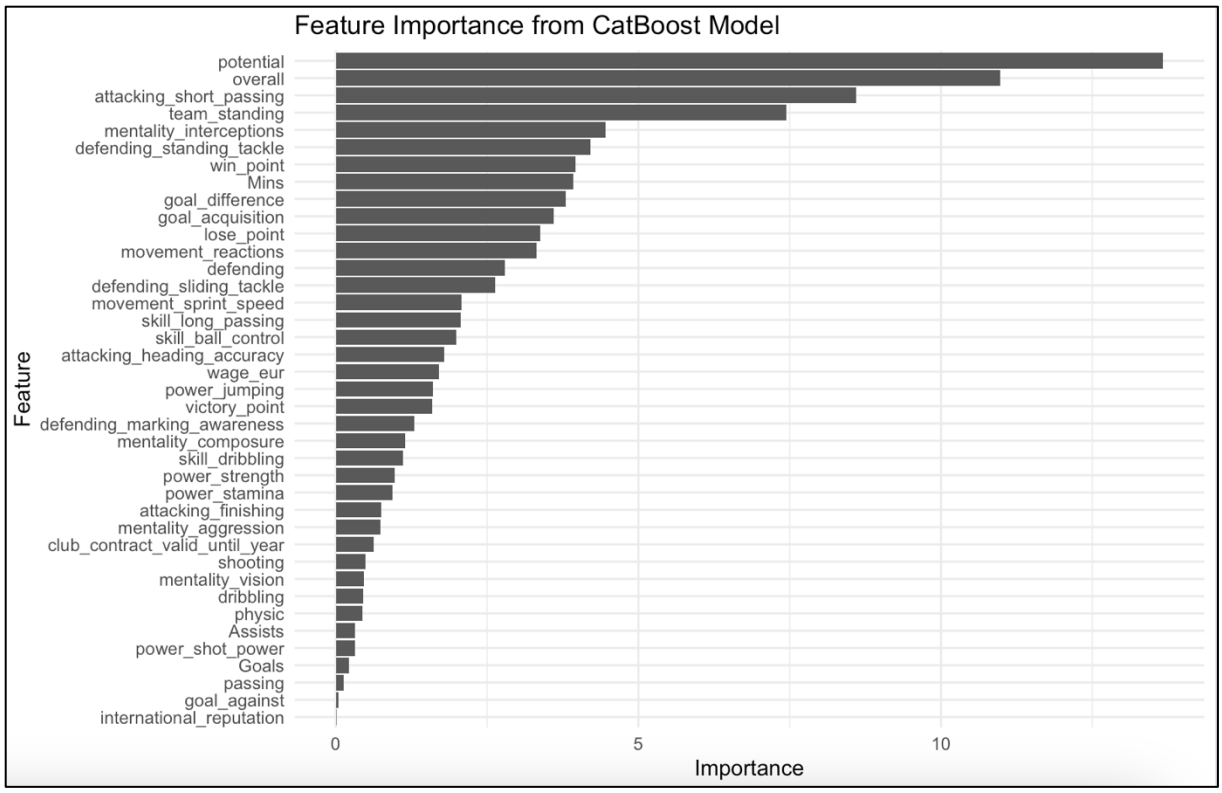

Figure 27: Feature selection and importance plot for CatBoost model with Bayesian Optimization for defenders

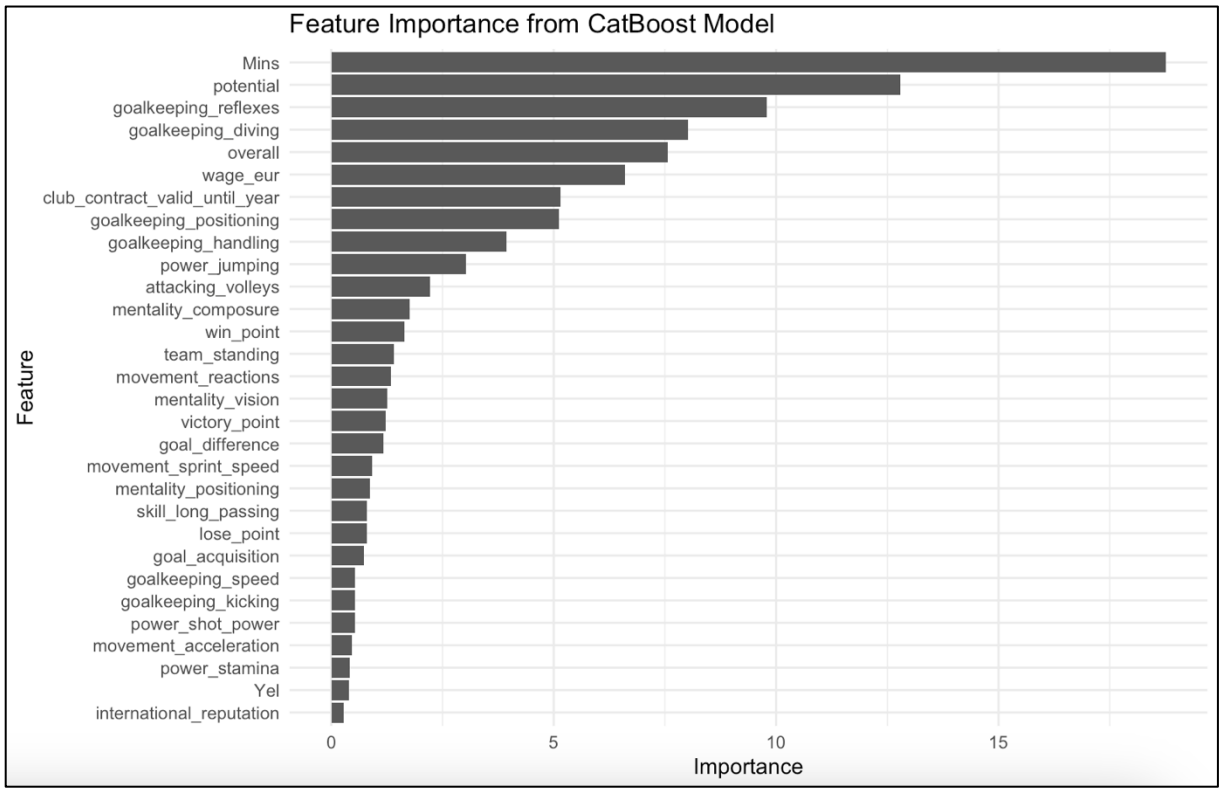

Figure 28: Feature selection and importance plot for CatBoost model with Bayesian Optimization for goalkeepers

Appendix 11: Performance in R2, MAE and F1-score for all models

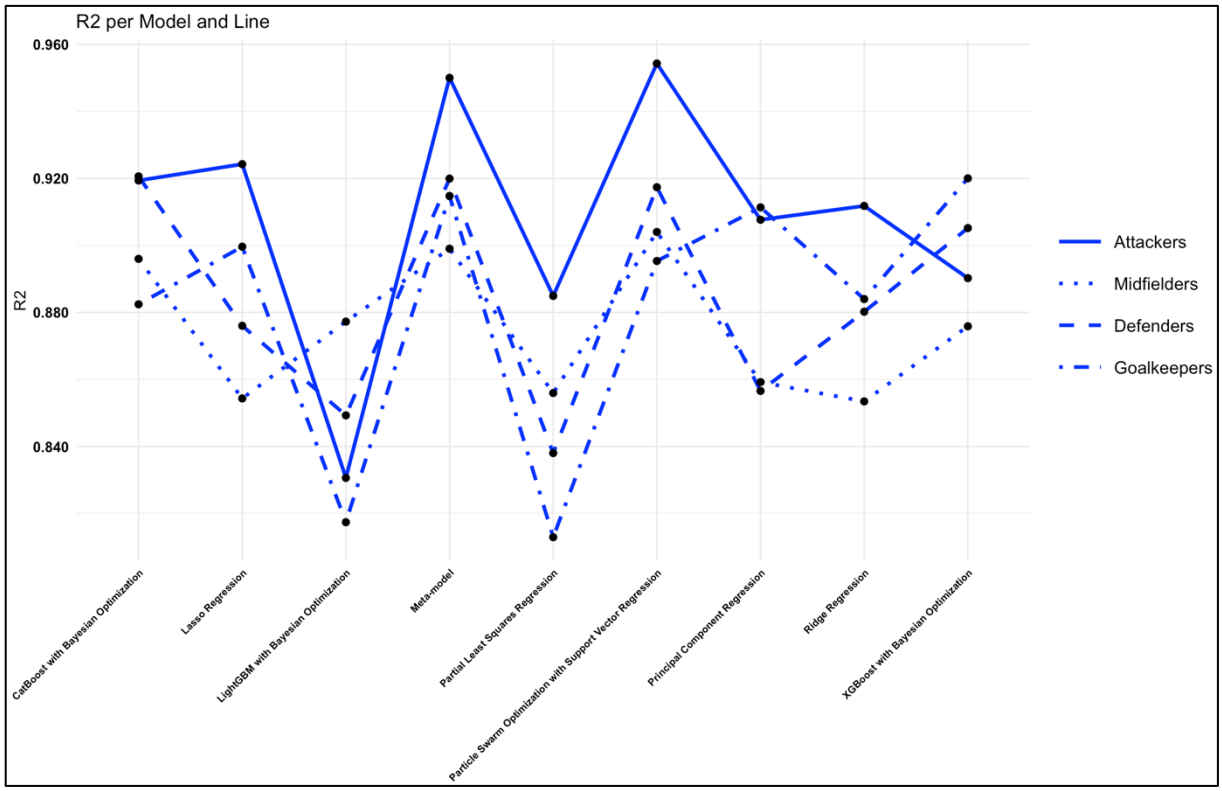

Figure 30: Performance in R2 per model

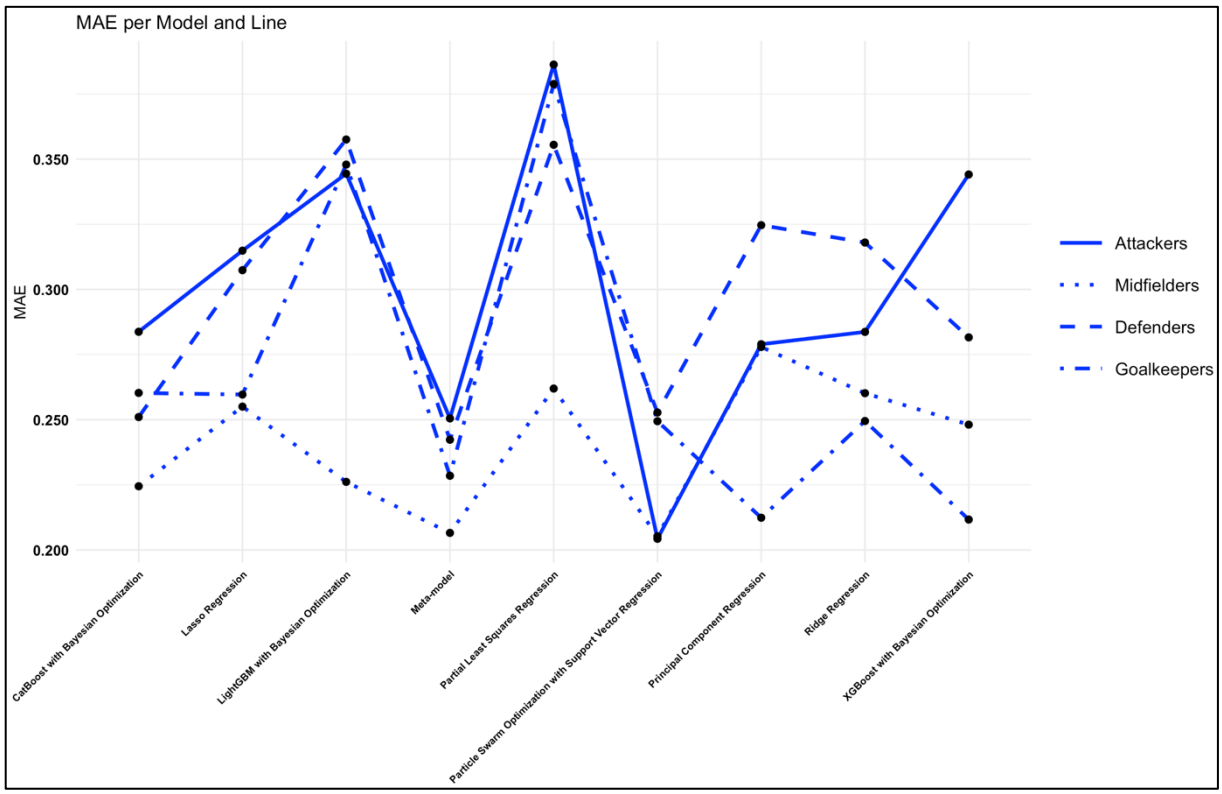

Figure 31: Performance in MAE per model

Appendix 12: R-script

Due to the length of the R-script that is used to facilitate all the data, preprocessing and model creation, a link to the file is given. Access to the file can be requested via the link below.

https://drive.google.com/file/d/1o64jA8IT8VoxS1HGQUF2wbVxxZx-hU4b/view?usp=share_link

References

Academic literature

Al-Asadi, M. A., & Tasdemır, S. (2022). Predict the value of football players using FIFA video game data and machine learning techniques. *IEEE access*, *10*, 22631-22645.

Arai, A., Ko, Y. J., & Ross, S. (2014). Branding athletes: Exploration and conceptualization of athlete brand image. *Sport Management Review*, *17*(2), 97-106.

Asif, R., Zaheer, M. T., Haque, S. I., & Hasan, M. A. (2016). Football (soccer) analytics: A case study on the availability and limitations of data for football analytics research. *International Journal of Computer Science and Information Security*, *14*(11), 516.

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, *25*(3), 2499-2511.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Breiman, L. (1996). Stacked regressions. *Machine learning*, *24*, 49-64.

Bryson, A., Frick, B., & Simmons, R. (2013). The returns to scarce talent: Footedness and player remuneration in European soccer. *Journal of sports economics*, *14*(6), 606-628.

Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: theory and evidence. *Applied Economics*, *25*(12), 1467-1476.

Chen, T., & Guestrin, C. (2016, 2016). Xgboost: A scalable tree boosting system.

Cotta, L., de Melo, P. O. V., Benevenuto, F., & Loureiro, A. A. (2016). Using fifa soccer video game data for soccer analytics.

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

Franceschi, M., Brocard, J. F., Follert, F., & Gouguet, J. J. (2023). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*.

Franck, E., & Nüesch, S. (2012). Talent and/or popularity: what does it take to be a superstar? *Economic Inquiry*, *50*(1), 202-216.

Frick, B. (2007). THE FOOTBALL PLAYERS'LABOR MARKET: EMPIRICAL EVIDENCE FROM THE MAJOR EUROPEAN LEAGUES. *Scottish Journal of Political Economy*, *54*(3), 422-446.

Fry, T. R. L., Galanos, G., & Posso, A. (2014). Let's get Messi? Top-scorer productivity in the European Champions League. *Scottish Journal of Political Economy*, *61*(3), 261-279.

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

He, M., Cachucho, R., & Knobbe, A. J. (2015, 2015). Football Player's Performance and Market Value.

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hofmann, J., Schnittka, O., Johnen, M., & Kottemann, P. (2021). Talent or popularity: What drives market value and brand image for human brands? *Journal of Business Research*, *124*, 748-758.

Inan, T., & Cavas, L. (2021). Estimation of market values of football players through artificial neural network: a model study from the turkish super league. *Applied Artificial Intelligence*, *35*(13), 1022-1042.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *31*(3), 300-303.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kennedy, J., & Eberhart, R. (1995, 1995). Particle swarm optimization.

Kiefer, S. (2012). *The impact of the Euro 2012 on popularity and market value of football players*.

Kumar, G. (2013). Machine learning for soccer analytics. *University of Leuven*.

Lee, H., Tama, B. A., & Cha, M. (2022). Prediction of Football Player Value using Bayesian Ensemble Approach. *arXiv preprint arXiv:2206.13246*.

Majewski, S. (2016). Identification of factors determining market value of the most valuable football players. *Central European Management Journal*, *24*(3), 91-104.

Medcalfe, S. (2008). English league transfer prices: is there a racial dimension? A re-examination with new data. *Applied Economics Letters*, *15*(11), 865-867.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, *263*(2), 611-624.

Pawlowski, T., Breuer, C., & Hovemann, A. (2010). Top clubs' performance and the competitive situation in European domestic football competitions. *Journal of sports economics*, *11*(2), 186-202.

Pollard, R., Reep, C., & Hartley, S. (2013). The quantitative comparison of playing styles in soccer. In *Science and Football (Routledge Revivals)* (pp. 309-315). Routledge.

Prasetio, D. (2016, 2016). Predicting football match results with logistic regression.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Puccio, G. (1999). Creative Problem Solving Preferences: Their Identification and Implications [https://doi.org/10.1111/1467-8691.00134]. *Creativity and Innovation Management*, *8*(3), 171-178. https://doi.org/https://doi.org/10.1111/1467-8691.00134

Sakınç, İ., Açıkalın, S., & Soygüden, A. (2017). Evaluation of the relationship between financial performance and sport success in European football.

Stanojevic, R., & Gyarmati, L. (2016). Towards data-driven football player assessment.

Supino, E., & Marano, M. (2024). Capital gains from player transfers as a value creation tool: some evidence from European listed football clubs. *Sport, Business and Management: An International Journal*, *14*(1), 80-98.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Szymanski, S., & Smith, R. (1997). The English football industry: profit, performance and industrial structure. *International review of applied economics*, *11*(1), 135-153.

Tenga, A. (2010). Reliability and validity of match performance analysis in soccer: a multidimensional qualitative evaluation of opponent interaction.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Vapnik, V. N., & Vapnik, V. (1998). Statistical learning theory.

Vroonen, R., Decroos, T., Van Haaren, J., & Davis, J. (2017). Predicting the potential of professional soccer players.

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44.

Wold, S., Ruhe, A., Wold, H., & Dunn, I. W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735-743.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241-259.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Yu, P. S. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1-37.

Yiğit, A. T., Samak, B., & Kaya, T. (2020). Football player value assessment using machine learning techniques.

Non-academic literature

Deloitte. (2023, June). Annual review of football finance 2023. Retrieved from www.deloitte.com: https://www2.deloitte.com/content/dam/Deloitte/de/Documents/consumer-business/Deloitte-Annual_Review_of_Football_Finance_2023.pdf

Football Manager. (2024). Sports Interactive.

KPMG. (n.d.). Player valuation. Retrieved from www.footballbenchmark.com: https://www.footballbenchmark.com/methodology/player_valuation

Scisports. (n.d.). Performance analysis. Retrieved from www.scisports.com: https://www.scisports.com/services/performance-analysis/