Beyond The Pitch Predictive and Explainable AI Applications in Football Analytics

Hans-Olav Skarpodde Goncalves

Dissertation written under the supervision of Professor Ana Marisa Mendes Gonçalves Vinhais Guedes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 02.01.2025.

Abstract

The application of machine learning in football analytics has significantly advanced, yet challenges remain in achieving a balance between predictive accuracy and interpretability. This thesis investigates the effectiveness of predictive models and explainable AI (XAI) techniques in forecasting football match outcomes and providing actionable insights for managerial decision-making. Historical match data, player statistics, and ELO ratings from the English Premier League (2017–2024) serve as the foundation for developing and evaluating machine learning models, including Random Forest, Gradient Boosting, and XGBoost.

Explainable AI techniques, such as SHAP (SHapley Additive exPlanations), are applied to interpret model outputs both globally and locally, revealing key predictors of match outcomes, including ELO differences, expected goals, and positional metrics. Formation simulations are utilized to assess the impact of various team setups on predicted outcomes, offering practical insights into tactical decision-making.

Results indicate that XGBoost achieves the highest predictive accuracy (55.2%), comparable to bookmaker odds provided by Bet365. SHAP visualizations enhance the interpretability of model outputs, identifying the features most influential in determining predictions and supporting more transparent decision-making processes. This research demonstrates the potential of combining predictive analytics with XAI to optimize tactical planning, improve player deployment, and refine strategic operations. By bridging the gap between complex predictive models and actionable insights, the study provides a robust framework for advancing data-driven decision-making in football analytics.

Title: Beyond the Pitch Predictive and Explainable AI Applications in Football Analytics

Author: Hans-Olav Skarpodde Goncalves

Keywords: Football Match Prediction, Machine Learning, XGBoost, Explainable AI

Resumo

A aplicação de machine learning na análise de futebol tem avançado significativamente, mas ainda enfrenta desafios em equilibrar precisão preditiva e interpretabilidade. Esta tese explora a eficácia de modelos preditivos e de técnicas de inteligência artificial explicável (XAI) na previsão de resultados de jogos de futebol e na geração de insights para a tomada de decisões de gestão. Utilizam-se dados históricos da Premier League Inglesa (2017–2024), abrangendo estatísticas de jogos, métricas de jogadores e classificações ELO, para desenvolver modelos de machine learning, incluindo Random Forest, Gradient Boosting e XGBoost.

As técnicas de XAI, como SHAP (SHapley Additive exPlanations), interpretam os resultados dos modelos global e localmente, identificando preditores-chave como diferenças de ELO, golos esperados e métricas posicionais. Simulações de formações são realizadas para avaliar o impacto de diferentes configurações de equipa nos resultados previstos, gerando insights práticos para decisões táticas.

Os resultados demonstram que o XGBoost alcança a maior precisão preditiva (55,2%), comparável às probabilidades da Bet365. Visualizações do SHAP aumentam a interpretabilidade, destacando as características mais influentes nas previsões e promovendo decisões mais transparentes. Este estudo comprova o potencial de integrar análises preditivas com XAI para otimizar o planeamento tático, aprimorar a gestão de jogadores e refinar operações estratégicas. Ao combinar modelos avançados com insights acionáveis, esta investigação fornece uma estrutura robusta para decisões baseadas em dados na análise de futebol.

Título: Para Além do Relvado: Aplicações Preditivas e Explicáveis de IA na Análise de Futebol

Autor: Hans-Olav Skarpodde Gonçalves

Palavras-chave: Previsão de Resultados de Futebol, Machine Learning, XGBoost, Inteligência Artificial Explicável

Table of Contents

Table of Tables

Table of Figures

1 Introduction

In recent years the application of machine learning has grown exponentially in sports analytics. Dating back to 1982 when M. J Maher used the Poisson Distribution to predict the number of goals in football games (Maher, 1982) to more sophisticated and complex models like Neural Networks, Random Forest, and Gradient Boosting models showing promising results (Arabzad et al., 2014; Baboota & Kaur, 2019). These advancements have continued to push the boundaries of predictive analytics in football.

Predicting football matches remains a difficult task (Arabzad et al., 2014; Spann & Skiera, 2009). This complexity arises from the dynamic nature of football, where outcomes are affected by numerous factors, such as player form, team tactics, injuries, and in-game calls made by managers. This is also exemplified in studies that show the accuracy in this specific prediction task, can be rather low (Arabzad et al., 2014). As mentioned, predicting football match results is an evolving field. Starting with the use of statistical models to enhance the accuracy of predictions (Maher, 1982). In more recent times machine learning models have been adapted to the field of sports predictions, to further increase the accuracy of predicting sports match outcomes (Bunker & Thabtah, 2019; Joseph et al., 2006; Prasetio & Harlili, 2016). While machine learning models outperform statistical models, the complexity of the task increases along with the accuracy, making it hard to know how the model came to its conclusion. Especially within the field of sports, much of the research is betting-related, and studies lack a focus on explainability (Baboota & Kaur, 2019; Bunker & Thabtah, 2019; Dixon & Coles, 1997).

Due to the increasing complexity of models a problem that is encountered is the "Black Box Problem" which essentially is that we do not know how a model came to its conclusion (von Eschenbach, 2021). These "Black Boxes" have been dealt with in different ways, where the field of model explainability is one of them. Using Explainable AI techniques we can draw insightful information about how models come to their conclusion.

This research aims to address these challenges by exploring the following questions:

(1) **How effectively can machine learning models predict English Premier League match outcomes using historic match, team, and player-specific data**

(2) **How can explainable AI be used to provide actionable insights for managerial decision-making based on player positional data?**

2 Literature Review

The literature review explores the evolution of predictive modeling in football and sports, highlighting the transition from statistical models to more sophisticated machine learning methods. Additionally, features affecting match outcomes, such as ELO ratings are examined as significant predictors. The studied literature also emphasizes the general use of Explainable AI and Interpretability, as well as how it has been applied to similar studies within sports prediction. Lastly, Table 2.1 provides an overview of studies and performance of previous models.

2.1 Data analytics in Football

One of the first authors to use statistical models for predicting football match outcomes was M. J. Maher in 1982, when he challenged other researchers who rejected the application of the Poisson model for football match predictions (Maher, 1982). Maher argued that the number of goals scored by a team is well approximated using a Poisson Distribution. Maher modeled each team's attacking and defensive strengths as independent variables, which revealed promising results in the field, showing that football match results can be explained by statistics, and not only by chance. This marked a shift in football analytics as it introduced a robust framework of goal predictions for future advancements in the field. Furthermore, this introduction laid the foundation for the research made by Dixon and Coles which extended Maher's work by introducing two main factors (Dixon & Coles, 1997). The Poisson Model developed by Maher assumed that the number of goals scored by the teams was independent. Dixon and Coles improved on this by implementing a bivariate Poisson Distribution that considers that there may be dependencies in goals scored by teams playing at home and away. The research also accounted for the model being too heavily weighted on historical data. By implementing a time-weighted parameter, the team's strengths and weaknesses were updated throughout the season. This introduced a better real-world case where teams can be in good and bad form. These modifications allowed for a more robust model with a better performance.

Another methodology for predicting football results was using the idea of Arban Elo (Elo, 1978), which developed a system for ranking chess players based on their performance. This system measured the performance of chess players by giving each player an initial rating. For example, if a higher ranked chess player wins against a lower ranked player, the rating of the winner will increase slightly. If the opposite happens the lower ranked players rating will

increase significantly. The ranking system developed by Elo showed promising results in predicting chess match outcomes with only 13 out of 114 matches ending in the lower-ranked player winning the match. This idea was then picked up on by Hvattum and Arntzen in their research of "Using ELO ratings for match result prediction in association football" (Hvattum & Arntzen, 2010). Hvattum & Arntzen used the methodology of Elo to create rankings of football teams in the English premier league. In the study, they used a Logit Regression and found out that using the ELO system for football was effective in predicting the outcome of premier league matches. Moreover, the authors used this system to aid them in predicting the outcome of premier league football matches and beating the bookmakers.

2.2 Machine learning in football

Machine learning is a field of study within Artificial Intelligence where models can learn from and make decisions based on data. It is efficiently used in the sports world today, especially by individuals trying to beat the bookmakers. In comparison with the statistical models from earlier years, machine learning models allow us to model more complex relationships than ELO Ratings and Poisson regressions (Sun, 2023).

The results within the field vary based on different studies and data. A study conducted by Joseph, Fenton, and Neil aimed at predicting the outcome of Tottenham's football matches, concluded that their KNN model performed well on the data from the same season, but did not generalise well on data from a different season (Joseph et al., 2006). This could indicate that the team form is an important aspect and explain why using the ELO system (Hvattum & Arntzen, 2010) performs well in predicting football results. The best model in the study was a Bayesian Network which reached 59.21% accuracy. A considerable limitation of this study was that the model was only trained to predict for one team. Jumping ahead in time Prasetio and Harlili achieved a 69,5% accuracy with a logistic regression predicting home and awaywins for 2015/2016 season (Prasetio & Harlili, 2016). The study achieved this using data from 2010/2011 until 2015/2016 using historic match data. Features in the study included both match-specific data such as home offence, and aways offence, as well as non-match-specific data, including travel fatigue from distance traveled, home advantage, and ground familiarity.

Nowadays, one of the most used approaches for sports performance prediction is based on deep learning, using Artificial Neural Networks (ANN) (Bunker & Thabtah, 2019). These models

have been shown to perform well in a study for the Iran Pro League (IPL), showing significant results using a ANN (Arabzad et al., 2014). In this study, the data comprised historic match metrics from the six last previous leagues including 29 weeks in the 2014 season. This resulted in 2068 records. The goal of the study was to predict the final game week of the IPL based on historic data. The study predicted the winners, and goals scored in the remaining 8 matches of the league. Predictions for the final game week were run 30 times to have a robust result. To evaluate the model ANOVA tests were used and revealed statistical significance for each match prediction, indicating the model's confidence in differentiating between the different team performances. X-MR charts were also applied to ensure prediction stability.

Another common approach is to perform comparison studies between multiple models. This is the case of Baboota & Kaur (Baboota & Kaur, 2019) reports, using historic match data collected for seasons from 2005 to 2016 where data from nine seasons was used for training and the remaining two for testing. The models tested were Naive Bayes, Linear SVM, RFB SVM, Random Forest and Gradient boosting. Naive Bayes showed the worst results with 51% and the best performing model was the Gradient Boosting model with an accuracy close to 59%. Though being lower than many other studies, it is important to state that the predictions were made for two seasons increasing the uncertainty. Feature importance is also addressed in the study using the Gini Index unveiling that the top features for predicting were Form Differential, Home Form, Away form, and Past number of shots on target.

2.3 Explainable Artificial Intelligence (xAI)

There are several methods that can aid in explaining and interpreting machine learning models. Models like Decision Trees, Linear Regressions and General additive models are called "White Box Models" due to their simplicity and their ability to clarify how features affect the model's conclusion (Dwivedi et al., 2023). As machine learning model accuracy has increased over time, the models have also become significantly more complex. This has led to less interpretable and explainable models. Algorithms are increasingly regulating people's lives. Decisions affecting human lives are dependent on machine learning models and the public have recently become more aware of this (Edwards et al., 2017). This introduces the "Black Box Problem". Eschenbach explains this problem as AI models that make decisions without understanding what makes the "Black Box" reach its conclusion (von Eschenbach, 2021). In the field of

Explainable and Interpretable Artificial Intelligence there is a need to open these "Black Box Models", and multiple methods have been developed within the field.
Models, such as Neural Networks and Random Forests require additional techniques for interpretation. Feature based methods, such as Partial Dependence Plots, Accumulated Local Effects and Individual Dependence plots can be applied to illustrate how specific features affect predictions either globally or locally to open these "Black Boxes". Model agnostic techniques can also give a better understanding of these models. An example of these methods are approximate black boxes with simpler interpretable ones. (Dwivedi et al., 2023).

A powerful methodology for explainability is SHAP, short for "SHapley Additive exPlanations" developed by Lee and Lundberg (Lundberg et al., 2017) is one of these. It is based on a game-theoretic approach where it assigns features an importance value for a particular prediction. SHAP provides both local and global insight to the model's predictions. A global insight is the average feature importance for all predictions made by the model, while a local insight allows you to dig into each prediction and see the feature importance for each observation. In addition to SHAP, another widely used method for explainability is LIME. LIME, short for Local Interpretable Model-agnostic Explanation explains feature importance locally for predictions (Ribeiro et al., 2016). Unlike SHAP using a game-theoretic approach, LIME approximates complex model predictions with simpler linear models. This is done by making perturbed samples and fitting these to a simpler model. This way it is possible to understand how variations affect the predicted output in an understandable way.

2.4 xAI in sports

Sports is affected by a multitude of factors as we have seen in the literature above. Furthermore, much of the research is betting-related, and does not take explainable AI into account. This makes the field an interesting topic to investigate to see if there are actionable insights that can be leveraged for decision-making within the field.
In a study on the National Basketball Association (NBA), researchers tried to predict gameplay leveraging explainable AI (Wang et al., 2022). The study used historical team, player, and game statistics to predict the outcome of NBA matches and leverage explainable AI to find what features affect the conclusions of the model. Models that were used in the study were a feed-forward neural network and a random forest classifier where the neural network performed the best. The top attributes in this study contributing to a positive value were found to be the "high

floor impact" (mean of impact each player gives), high defensive rating, and good offensive power. The study concludes that these features make sense from a basketball standpoint and follow general coaching principles. Interestingly, the features that contributed the most to inhibiting a high win ratio were a high number of games played, a high number of steals, and a high number of assist percentages. Similarly, another study focusing on 11 top European football leagues tried to leverage SHAP to gather insightful information on what factors affect team performance (Moustakidis et al., 2023). The researchers in this study aim to predict the average goal difference in a season based on team, and player statistics throughout the season. The predictions were done using Support Vector Regression, Random Forest, k-nearest-neighbor Regressor, and XGBoost, where the latter outperformed the others with a root mean squared error of 32,09%. The study leveraged SHAP to get a local insight to the individual predictions, and interestingly found that the top features for predicting varied from team to team. An average showed that some of the most important features for predicting were "shots per quantity possession percent", "missed chances" and "entries to the penalty box". Even though these features are the global average, the study also revealed that the feature importance varied from team to team.

To provide a comprehensive overview of key studies from relevant for this study, Table 2.1 summarizes main characteristics, methodologies and performance. The table highlights scope of the previous research, including datasets used, specific focus, and methods applied. By summarizing this information, the table also show trends and gaps in the field of sports and machine learning.

Table 2.1 - ***Overview of studies using machine learning to understand football.*** *For each study the main models used are listed, as well as the key details about the followed methodology and performance metrics.*

Study	Model used	Comments	Performance
(Joseph et al., 2006)	Bayesian Network	Predictions were specific for one team	59,21 % accuracy
(Prasetio & Harlili, 2016)	Logistic Regression	Used previous research to only include proven variables that work. Also included video game data (FIFA). Did not account for draws	69,5 % accuracy
(Arabzad et al., 2014)	Artificial Neural Network	Predicting the exact score in the Iranian	ANOVA showed statistical

		Pro League. Only for the last 8 matches of the league.	significance for each prediction. X-MR Charts ensured prediction stability
(Baboota & Kaur, 2019)	Gradient Boosting	Predictions were made for 2 seasons into the future	59% Accuracy
(Wang et al., 2022)	Artificial Neural Network	Predictions were made on the NBA. Leverages LIME for Explainability	R Squared of 0,77 RMSE of 0.009
(Moustakidis et al., 2023)	XGBoost	Predicts team performance. Leverages Shap for explainability	RMSE of 0,3209

3 Methodology

This thesis examines the effectiveness of machine learning in predicting Premier League match outcomes, utilizing historical data on matches, teams, and players. By integrating explainable AI techniques, it also seeks to derive actionable insights to inform managerial decision-making.

This methodology chapter outlines the research process, detailing data collection, cleaning, feature engineering, model evaluation, and explainability approaches. These steps were designed to ensure methodological rigor, minimize bias, optimize model performance, and achieve the study's objectives. Figure 3.1 provides a summary of the applied methodology.

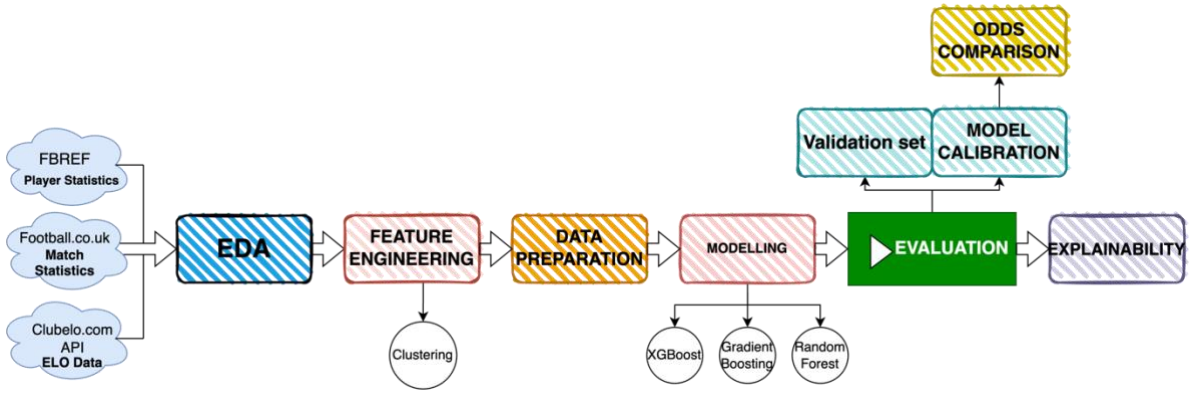

Figure 3.1 ***Overview of studies using machine learning to understand football.*** *For each study the main models used are listed, as well as the key details about the followed methodology and performance metrics.*

3.1 Data collection

The data collection done in this study was based on previous studies and performance within the field of football match prediction (Arabzad et al., 2014; Baboota & Kaur, 2019; Bunker & Thabtah, 2019; Hvattum & Arntzen, 2010; Joseph et al., 2006; Maher, 1982; Prasetio & Harlili, 2016). Combining this knowledge with data availability resulted in three datasets: i) historic match data; ii) historic player statistics; and iii) historic ELO ratings of all premier league football clubs.

3.1.1 Match Data

Match level data was collected from football-data-co.uk (Football-Data, n.d.), a football betting, odds, and results service. This database holds data from multiple football leagues and for this study data from the English Premier League for the seven seasons (2017/2018 – 2023/2024) was downloaded.

The reasoning behind the choice of these seasons was due to new features being introduced in the 2017/2018 season, improving data consistency and allowed for easier joining of the tables, with minimal missing values. This ensures higher-quality data for analysis and predictions. This historic match data consists of 28 features such as match outcome, date, match statistics and betting odds (see Table 7.1 in appendix for a detailed description of all variables as a data dictionary).

3.1.2 Player Data

Player level data was collected via web scraping on the football statistics library FBREF (FBREF, n.d.). FBREF's data uses combined historic data with match statistics provided from Opta (Statsperform, n.d.). Opta is a British sports statistics company and a world leader in Sports AI and the official English Premier League data provider (premierleague.com, n.d.). The acquired dataset was scraped for the same seasons as the historic match data (17/18 to 23/24 season).

The player-specific dataset contains one observation per player per season, which is more aggregated than historical data. The dataset consists of 3687 rows, and 38 columns. The data is player-specific for each season with features like age, squad, minutes played, and more in-depth on-field features (see Table 7.2 in appendix for mode detailed description of all variables).

3.1.3 ELO Data

ELO data was obtained from clubelo.com (Lars Schiefler, n.d.)., using the official API to scrape data for all teams that have been in the premier league storing them in a dataset. Clubelo provides historic club ELO data for European football clubs with daily updates assuring the dataset is up to date. The Data from 1946 to 2024 was collected and contains 209139 rows and 8 features, such as date, ELO ranking, and club name (see table 7.3 in Appendix for full feature description).

3.2 Data preparation

3.2.1 Data Wrangling

The exploratory data analysis began with analyzing the data in its raw format. During the EDA the following was identified: relationships, data structure, missing data, and wrongly formatted data. Distributions and outliers were visualized, and the target distribution was checked for imbalance. As there were three outcomes, the distribution of these were analyzed.

3.2.2 Joining the datasets

After the EDA, ELO and Historic match data was joined together for further work. An issue that could be encountered was data leaks, due to the temporal dependencies here. Because of this, a measure that makes sure the ELO ratings are joined 1 day before the match happens was added, making sure the model did not learn something from an increase or decrease on the match day.

The player statistics data was not joined with the rest of the data yet. This is due to the different levels of the data. As the player statistics data is seasonal, and the historic match data was match-based they could not be joined in the current format. The features engineered from the player-specific dataset were designed to address our research questions, especially by incorporating positional information to aid decision-making. Keeping this in mind, there was created multiple positional and team-based features. To avoid these features leaking information to the model, the player statistics data was lagged one season backwards before joining with the match statistics dataset.

3.2.3 Missing data and outliers

After combining datasets from several sources, missing data was analyzed, including the loss of a full season due to the inclusion of lagged data. The first season was omitted rather than imputed to avoid bias and preserve data variability. Remaining missing values were handled using scikit-learn's SimpleImputer with the median strategy, applying the median calculated from the training set to impute missing values in the test and validation sets. This approach preserved columns with missing data without significant information loss.

3.2.4 Feature Engineering and Codification

Having all the information combined in one panel of data, new features were created, aligned with what has been described in the literature review. The dataset features were created to capture team performance, squad composition and player attributes for both home- and awaywins. Team Form was represented through recent 3-game win streaks, and the difference in recent form (form_diff) took wins and losses of the opposing teams and summed the difference. Team strength was further quantified through ELO difference (elo_diff) which accounted for the difference in ELO between teams. Squad specific attributes included number of players in each positition (e.g midfielders, forwards and defenders) for both starting lineups

and the entire squad, as well as the average age in of players in each position. The positional data was used to create formations for each team, where the eleven players with highest number of started matches were used. Expected goals reflected the offensive potential of the teams and is a sum of the individual Expected Goals for each player grouped by team and season. Additionally, all categorical variables were converted into dummy variables. K-means clustering was also applied to incorporate cluster-based player distributions to identify patterns in player types. This is explained in the following subsection. See table 7.4 in appendix for a full overview of engineered features,

3.2.4.1 K-Means Clustering

To be able to draw more information about the players in the dataset K-Means Clustering was applied. K-Means Clustering is an unsupervised machine learning technique that groups data into clusters based on their similarities. (Ahmed et al., 2020). The number of clusters ("k") is how many different clusters the data is split into. There are multiple ways of choosing number of clusters, where both domain knowledge and statistical methods can be applied (Kodinariya & Makwana, 2013). Due to the nature of the data, and task of this study within explainability, 11 clusters were chosen to be able to classify the different player types where 11 different positions are used in a football match. Some players in the dataset have multiple positions, such as both Midfielder and Forward. As a result, analysts and stakeholders can better understand player types and tactical roles, ultimately improving the interpretability.

3.2.4.2 Principal Component Analysis (PCA)

PCA serves as a dimensionality reduction technique, transforming high dimensional data into a lower dimensional space, while retaining the most critical information by identifying the largest variance (Ding & He, 2004). Furthermore, the data was reduced to a two-dimensional dataset, which then was analyzed through a scatterplot, allowing for analysis and distinguishing clusters from each other.

The clusters created from player statistics data was then aggregated on a team level, which allowed for insertion of number of players in each cluster for each team. Both for the total and the starting 11. These clusters were used as inputs for the machine learning models to improve accuracy as well as explainability.

3.2.5 Train test split

To prevent data leakage and guarantee similar predictions over seasons, even with the decrease in data, the dataset was subsequently separated into training (seasons 2–5), testing (season 6), and validation (season 7) sets.

3.3 Modelling

This section details the process of building and optimizing the machine-learning model used in this study. Following the data preprocessing, the next step was to implement a suitable model to predict the outcome of matches. For the prediction task the target variable was Full Time Result (FTR). FTR could result in three different outcomes; HomeWin, Draw or AwayWin. A set of different models were used in the prediction task. The chosen models were Random Forest, XGBoost, and Gradient Boosting. Multiple models were chosen due to modern toolkits allowing us to easily compare different models, and their strengths in catching complex, non-linear relationships in the data (Bentéjac et al., 2021; Provost & Fawcett, 2013).

3.3.1 Random Forest

Random forest is one of the most successful machine learning models (Bentéjac et al., 2021; Breiman, 2001). The algorithm is an ensemble of decision trees and is designed to improve predictive accuracy and control for overfitting by introducing randomness in two ways. The randomness is generated by taking random samples of data, as well as random selection of features (Bentéjac et al., 2021).

3.3.2 Gradient Boosting Models

Gradient Boosting models, like the XGBoost and Gradient Boosting are models used both for classification and regression. The models combine predictions of weak learners to form a more accurate ensemble model. The core idea is to minimize the loss function through gradient descent (Bentéjac et al., 2021). The Gradient Boosting implementation is a general implementation of the algorithm, while the XGBoost model is an efficient and scalable implementation of Gradient Boosting. It also introduces regularization to prevent overfitting, and optimization such as parallel processing.

3.4 Model Tuning

The modeling process includes setting up initial parameters, training and validation process, and application of hyperparameter tuning. The hyperparameters for each model were tuned

using grid search method and 10-fold cross-validation to identify the optimal configuration. Grids searched can be found in table 7.6 in appendix. Table 3.1 provides an overview of the resulting parameters. For Random Forest the depth of trees, minimum number of samples required to split a node, and the number of trees in the ensemble were configured to improve predictions. Gradient Boosting and XGBoost were tuned for the learning rate, tree depth, and number of boosting iterations to achieve a tradeoff between learning efficiency and overfitting. Table 3.2 explains hyperparameters for each model in depth.

Table 3.1 - **Chosen Hyperparameters Based on Tuning.** Overview of selected hyperparameter values for Random Forest, Gradient Boosting, and XGBoost models after optimization.

Model	Parameters
Random Forest	Max depth: 10, min samples split: 2, n estimators: 200
Gradient Boosting	Learning rate: 0.01, max depth: 3, n estimators: 100
XGBoost	learning_rate: 0.01, max depth: 3, n estimators: 200

Table 3.2 - **Overview of hyperparameters and what they do.** Overview of hyperparameter abilities for Random Forest and Gradient Boosting models. Gradient Boosting and XGBoost hyperparameters are the same.

Model	**Hyperparameter**	**What It Does?**
Random Forest	Max Depth	Limits the maximum depth of each decision tree to control overfitting and tree complexity.
	Min Samples Split	The minimum number of samples required to split a node. Reduces unnecessary splits when set higher.
	N Estimators	Number of decision trees in the ensemble. More trees improve accuracy but increase training time.
Gradient Boosting Models	Learning Rate	Controls the step size for each boosting iteration, preventing overfitting when set low.
	Max Depth	Limits the depth of trees, reducing model complexity to avoid overfitting.

	N Estimators	Total number of boosting iterations (trees). A higher number compensates for a low learning rate.

3.5 Model Evaluation

Model performance was assessed using a diverse set of performance metrics. Primary metrics include Receiver Operating Characteristic (ROC) curve, Confusion Matrix and matrix-derived stats, such as Accuracy, Precision, Recall and F1 Score.

1. **ROC Curve and AUC Score**

   The ROC Curve provides a graphical representation of the model's performance across different classification thresholds. It shows the True Positive Rate versus the False Positive Rate. This allows us to evaluate how well the models distinguish between classes. The Area Under the Curve metric is used as a summary. The higher the AUC the better the model is performing.

2. **Confusion Matrix**

   The Confusion Matrix breaks down the model's predictions into a matrix. This matrix holds the values for True Positives, True Negatives, False Positives, and False Negatives. From this matrix, several key metrics are derived.

   Accuracy measures the total proportion of correct predictions done by the model

   Precision indicates the proportion of positive predictions that are correct.

   Recall (Precision) Shows the model's ability to classify true positives

   F1 Score Combines recall and precision into a single metric and is useful when data is imbalanced as it balances both false positives and false negatives

Evaluating the models on these metrics ensures that the models accurately and reliably could capture nuances of the data while minimizing errors.

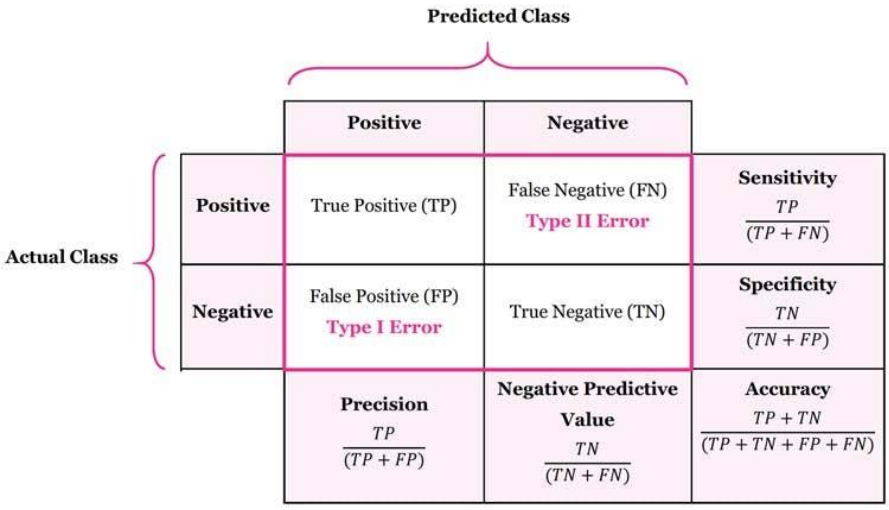

Figure 3.2 - **Confusion Matrix: Metrics and Error Analysis.** Visual representation of the confusion matrix, detailing classification metrics such as sensitivity, specificity, precision, and accuracy, alongside Type I and Type II errors.

3.5.1 Odds comparison

To further evaluate the results the Bet 365 odds data in our dataset was used, which is way to be able to compare the results. For this analysis the validation dataset was used. Bet 365 provided the odds for “HomeWin” (H) , “Draw” (D) and “AwayWin” (A). As the model is a multi-class model and also outputted probabilities it was possible to compare these to see if they are performing worse, better or on par with the odds. Firstly, the odds needed to be converted to probabilities (Grimes & Schulz, 2008). This was done using the following formulas in figure 3.3:

$$\text{prob}_{\text{b365h}} = \frac{1}{\text{b365h}}, \quad \text{prob}_{\text{b365d}} = \frac{1}{\text{b365d}}, \quad \text{prob}_{\text{b365a}} = \frac{1}{\text{b365a}}$$

Figure 3.3 - ***Formulas for Converting Betting Odds into Probabilities.*** *Mathematical conversion of Bet365 odds for "Home Win" (H), "Draw" (D), and "Away Win" (A) into corresponding probabilities, enabling direct comparison with model predictions.*

To ensure that these probabilities summed up to 1 they needed to be normalized by dividing each initial probability by the total sum of all three probabilities. This normalization ensured that all probabilities sum to 1, representing the bookmaker's implied probability for each outcome. The formulas applied are explained in figure 3.4.

$$\text{prob}_{\text{b365h}} = \frac{\text{prob}_{\text{b365h}}}{\text{prob}_{\text{b365h}} + \text{prob}_{\text{b365d}} + \text{prob}_{\text{b365a}}}$$

$$\text{prob}_{\text{b365d}} = \frac{\text{prob}_{\text{b365d}}}{\text{prob}_{\text{b365h}} + \text{prob}_{\text{b365d}} + \text{prob}_{\text{b365a}}}$$

$$\text{prob}_{\text{b365a}} = \frac{\text{prob}_{\text{b365a}}}{\text{prob}_{\text{b365h}} + \text{prob}_{\text{b365d}} + \text{prob}_{\text{b365a}}}$$

*Figure 3.4 - **Normalizing Probabilities.** Formulas for normalizing Bet365 probabilities to ensure they sum to 1, representing the bookmaker's implied probability distribution for each outcome*

3.6 Model Calibration

In machine learning classification problems, calibrating probabilities is essential when the predicted probabilities are not well-calibrated, meaning they do not accurately reflect the true likelihood of an event. To calibrate the probabilities a Logistic Regression model was fitted to the predicted probabilities using a technique called Platt scaling (Böken, 2021). This scaling adjusts the raw output probabilities of the model by fitting a Logistic Regression to the raw probabilities of the model, scaling them to calibrated probabilities that better represent true likelihoods. With the now calibrated probabilities, the model's predictions could be compared with the odds, and compute metrics like RMSE, MAE and Confusion Matrixes.

3.7 Model Explainability

"Black box" models due to their complexity, need additional techniques to allow for interpretation of variable importance compared to "White box" models. A method for this is SHapley Additive exPlanations (SHAP). SHAP is a powerful tool for model interpretability and provides a unified approach to explain individual predictions based on game theory (Lundberg et al., 2017).

SHAP values are an approach using a game theoretic approach which distribute the predictions among features in a "fair" way. The approach makes it a robust method for feature importance and ensures that the models features can be attributed to each prediction giving insights to how the model came to its conclusion.

1. **Global Interpretation**

   SHAP allows for both global local interpretability. At the global level, SHAP values provide an overall feature importance, indicating which features have the most significant impact on the models' predictions. This can be visualized through summary plots which show the distribution of the feature importance for the overall predictions of the model. These insights help to identify key features for predictions, making it possible to identify the key factors for the outcome and understand the model's behavior.

2. **Local Interpretation**

   In addition to global interpretation, SHAP offers the possibility for Local Interpretation. Local interpretation contrasts with Global, the ability to understand specific predictions. This allows us to understand the underlying factors of how a feature positively or negatively affects one specific prediction. This level of detail is valuable in practical scenarios as it allows us to explain specific predictions to end-users and stakeholders.

3.8 Data Simulations

To further explore the potential outcome of matches and how features affect predictions, simulations of multiple scenarios were performed. As this study investigates how managerial decision-making material can be drawn using explainable AI techniques, different formation types are being investigated. Using the validation set, simulations were ran to see how the model's predictions change when different formations were pitted against each other. The reasoning behind using the validation set is trying to make this a real-world scenario where the data is unseen by the model. Using computational power, thousands of different scenarios for each single prediction was generated, allowing the testing of multiple predictions, and can then outline the predicted best formations on a global and local basis.

4 Results

4.1 Exploratory Data Analysis (EDA)

The datasets analyzed consisted of multiple match datasets, where the main dataset contains historic match data from the premier league. Figure 4.1 a shows target balance in the dataset and use Full Time Result (FTR) as the target variable. FTR is split into either H, D, or A which represent HomeWin, draw and AwayWin. There was one observation per match which gave 380 matches per season, and a total dataset of 2660 rows over 7 seasons. The data was distributed with 44,8% HomeWins, 32,6% AwayWins, and 22,6% draws (Figure 4.1 left) illustrating a small target imbalance, and consisted of 30 unique teams (Figure 4.1 right). Some teams had more matches than others due to relegations and promotions. The dataset had no missing values.

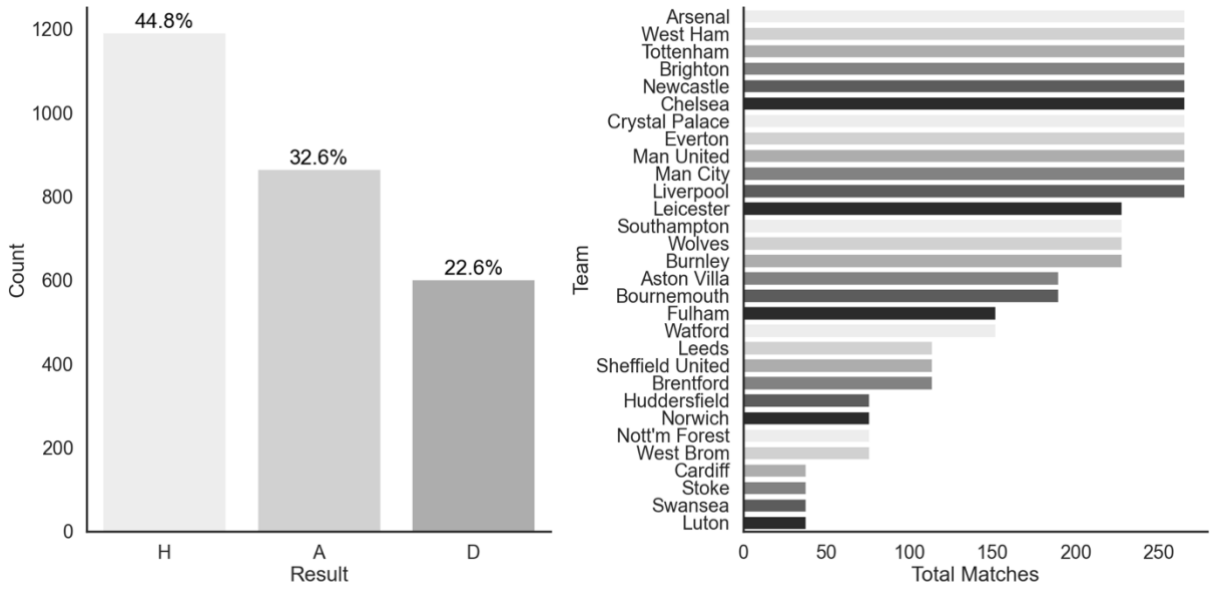

Figure 4.1 - **Distributions of Target and Teams**. Left: Distribution of target result (H = HomeWin, D = Draw, A = AwayWin). Right: Distribution of total matches per team

4.1.1 Correlation Matrix of match-specific features

The correlation matrix (4.2 left) showed correlations between all variables including the target variable. Looking at the matrix the highest correlated variables were identified as the betting variables Bet 365 Draw (b365d) and Bet 365 Away Win (b365a). The target showed strong positive and negative correlation with Full Time Home Goal (fthg) and Full Time Away Goal (ftag) showing values of 0,65 and -0,65.

4.1.2 Implied probabilities and Actual outcomes

Figure 4.2 (right) showed the comparison of the implied probabilities made by the betting odds, compared to the actual outcomes. This comparison illustrated how well the odds predicted the outcome of football matches for the English premier league with little over- and underestimation of the probability of outcomes compared to the actual outcomes, making them a good metric of comparison.

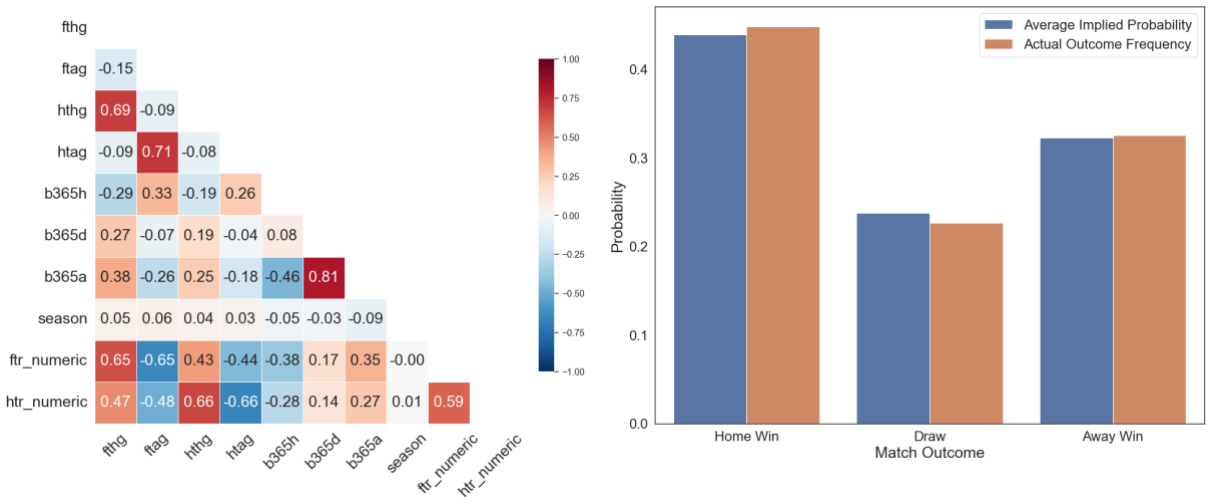

Figure 4.2 - ***Match Correlation Matrix and Implied vs Actual probability.*** *Left: Correlation Matrix, Right: Implied vs Actual probability*

4.1.3 Temporal dependencies

As the dataset was temporal it held temporal dependencies. Temporal dependencies are relationships or patterns in the data which are influenced by time. The temporal dependencies were handeled by lagging static seasonal data, such as the player statistics data by one season. For the ELO data it was made sure that ELO was not updated until after the match had happened. This is explained in the methodology section 3.2 and section 3.3. The results show that the most common outcomes were HomeWin, AwayWin, and Draw (Figure 4.3). Looking at this over time, it was mostly consistent with one exception of season 4 where there were more away wins than homewins. Furthermore, draws was the most consistent outcome in where the proportion was mostly the same except for in season 1.

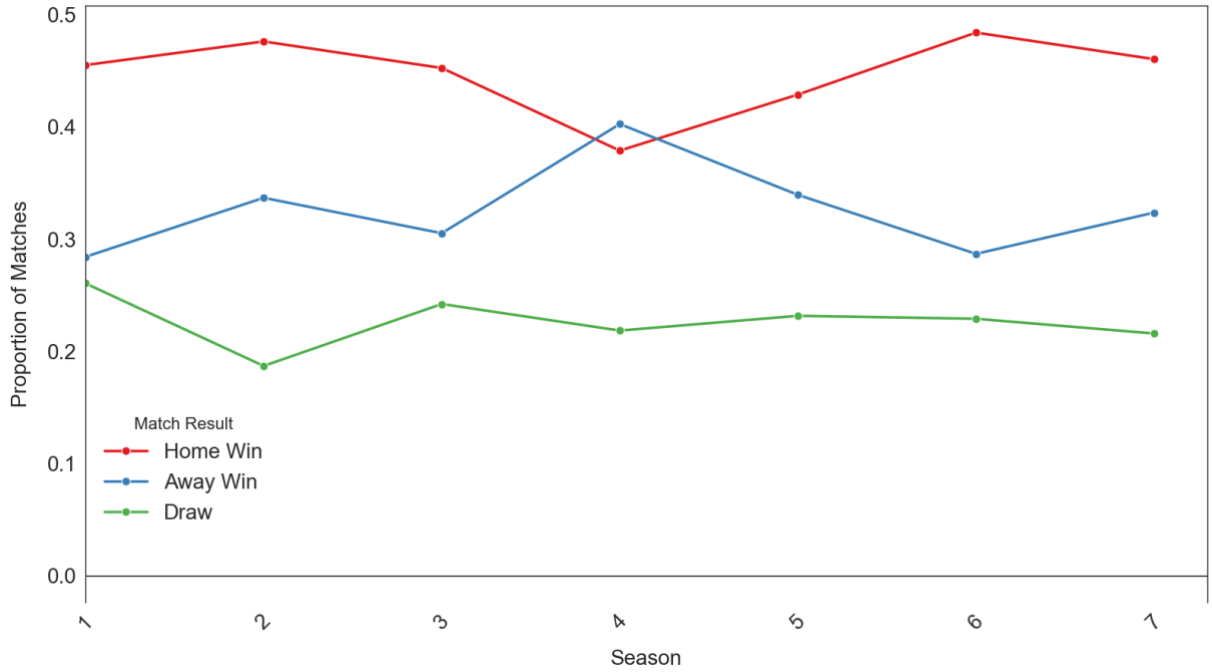

*Figure 4.3 - **Time Series Analysis of Target Variable Distribution.** Proportion of match outcomes (Home Win, Away Win, and Draw) across seasons, highlighting trends in target variable distribution over time.*

4.1.4 Club ELO

The Club ELO dataset consisted of the ELO ranking of all teams dating back to 1946 and showed the development of the team rankings (Figure 4.3). In this case, only the data in the seasons 17/18 to 23/24 was used and was joined to the data based on the date, giving insights on team form and performance. It was made sure that the ELO data was only joined before the match happened to avoid any data leaks. To illustrate ELO's development, the top three teams were shown in figure 4.4. This was based on the current top teams and shows Chelsea, Liverpool, and Manchester City.

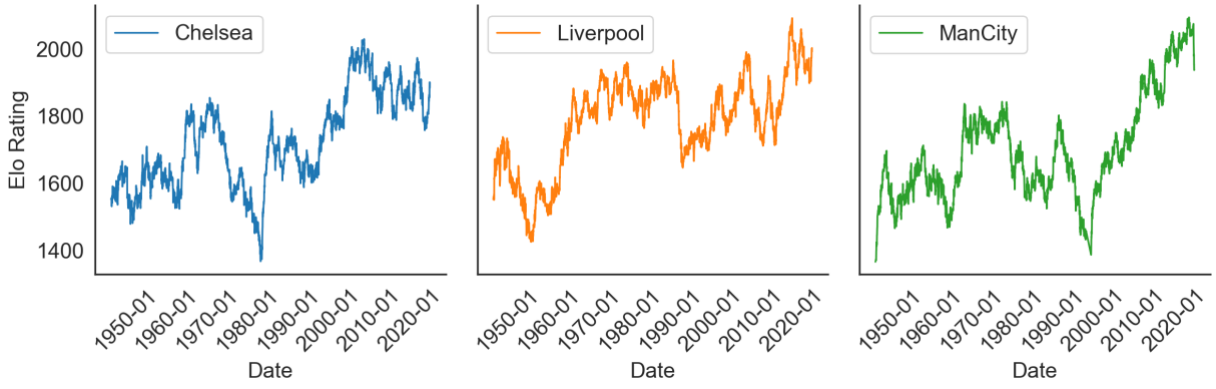

*Figure 4.4 - **Time Series Analysis of ELO Ratings for Top 3 Ranked Teams.** ELO rating trends over time for Chelsea, Liverpool, and Manchester City, showcasing the historical performance of these top-ranked teams.*

4.1.5 Player Statistics

The correlation heatmap provides insights into relationships between various performance metrics among players (Figure 4.4 left). Strong positive correlations are identified between metrics like Goals (Gls) and Goals + Assists (G+A), Expected Goals (xG) and Expected Assists (xAG), and Progressive Carries (PrgC) and Progressive Receives (PrgR). Moreover, there was certain metrics with weak and near zero correlations, such as Goals (Gls) and Yellow Cards (CrdY), among other slightly negative correlations with Yellow Cards across different metrics. Furthermore, in Figure 4.5 (right) Average Progressive Actions by position was analyzed. Forwards had high average values in Progressive Receives, midfielders showed high values in progressive passes, defensive players showed more balanced values, having a lower average value in most of the variables. Notably the goalkeeper illustrated close to zero values.

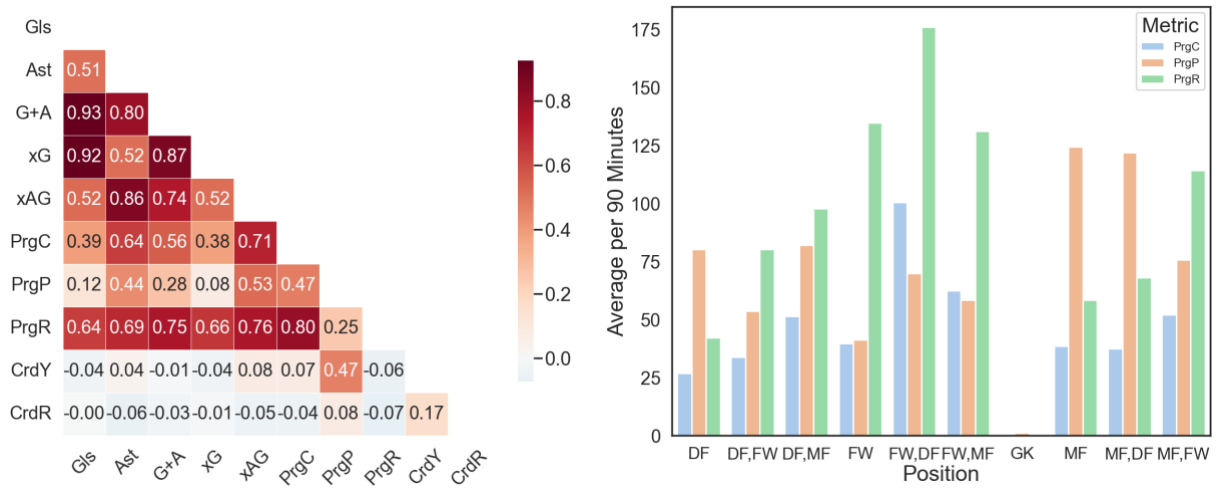

Figure 4.5 - ***Correlation Matrix of Player Statistics and Distribution of Progressive Metrics by Position.*** *Left: Correlation matrix of key player performance metrics. Right: Average values of progressive metrics (PrgC, PrgP, PrgR) across player positions, highlighting positional tendencies in contributions.*

4.1.6 Feature Engineering

Some of the most significant features developed during the feature engineering process are highlighted in this section. Important dynamics and player-specific characteristics were recorded to enhance forecast accuracy by concentrating on elements like team configurations and player grouping. Some noteworthy, engineered characteristics and how they related to the model's performance where shown below.

4.1.6.1 Team Form

Looking at the heatmap in figure 4.6 top performing teams were identified, where Manchester City and Liverpool were identified as the most consistent teams. In contrast, teams struggling were identified like Fulham, Huddersfield, and Sheffield United, where Fulham showed a record low in season 2. Some teams, such as Arsenal and Chelsea, showed fluctuating performances over the seasons with highs up to 13,7 and lows to -0,3, where a minus value means that more matches were lost than won. Several cells also had a value of 0, which indicated seasons where teams did not participate in the English Premier League due to relegation. Teams like Cardiff, Leeds and Norwich have multiple zero values across seasons, which gave less training data on these teams.

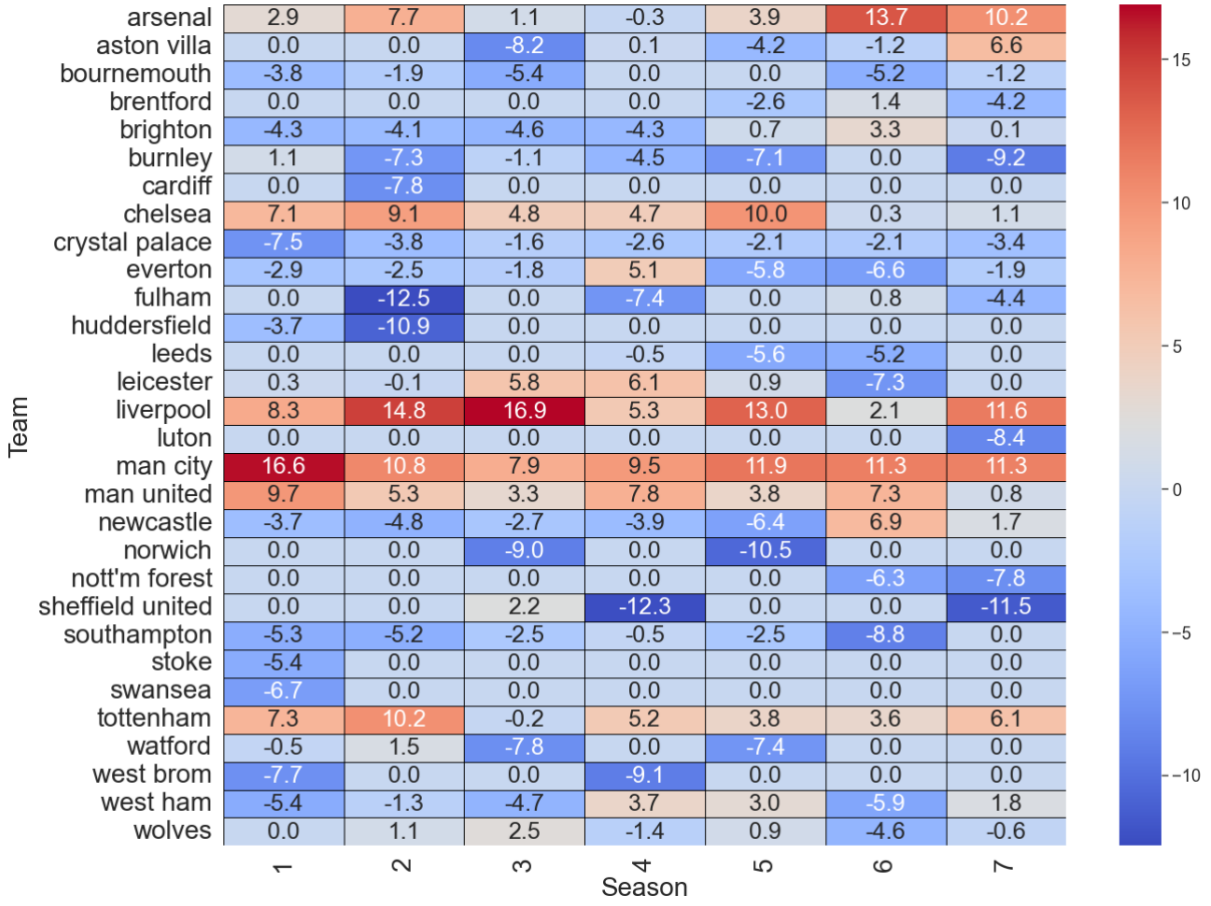

*Figure 4.6 - **Development of Team Form Per Season.** Heatmap showing the variation in team performance across seasons, with positive values indicating more wins than losses and negative values representing more losses than wins. Zero values denote seasons where teams had equal wins and losses or did not participate.*

4.1.6.2 Player Clusters

The Principal Component Analysis (PCA) in Figure 4.7 of player clusters revealed distinct groupings of players based on their characteristics. Clusters 4, 5 and 6 contain a high density of players positioned close to the center, suggesting players with common attributes. Some clusters such as 9 and 10 were located farther from the center and exhibit fewer players. These clusters show outliers and likely represented outliers or players with unique skills that differentiate from the rest. The distance to the other clusters suggested that these players had players that differentiate from the rest in the dataset. Looking at the whole visualization it showed that there are some player types with similar characteristics, while some had more outliers and length from the center indicating unique player styles.

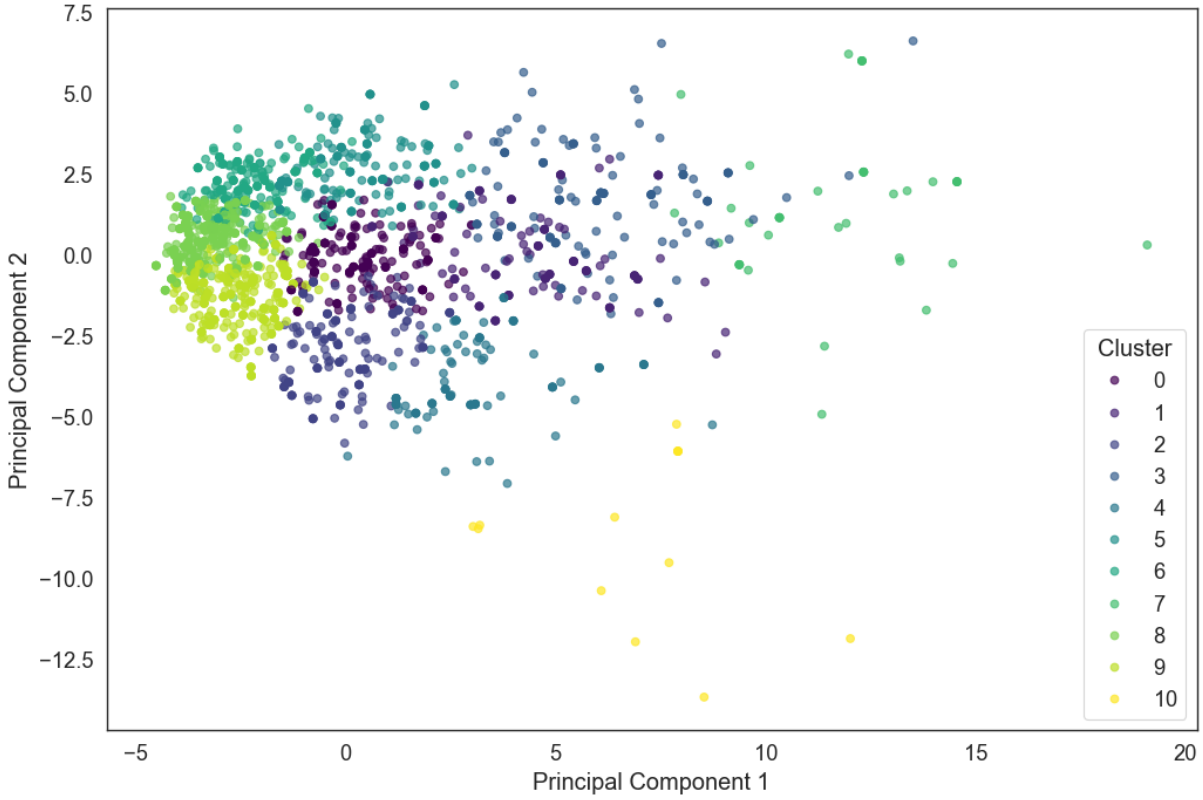

*Figure 4.7 - **Principal Component Analysis (PCA) of Player Clusters**. Scatterplot of player clusters based on PCA, illustrating distinct groupings of players with similar characteristics across principal components.*

Figure 4.8 of the elbow method suggested using between four and five clusters, domain knowledge and practical considerations led to selecting 11 clusters. This decision not only allowed for a more granular classification of player types but also improved accuracy by ~4% compared to using five clusters on the test set.

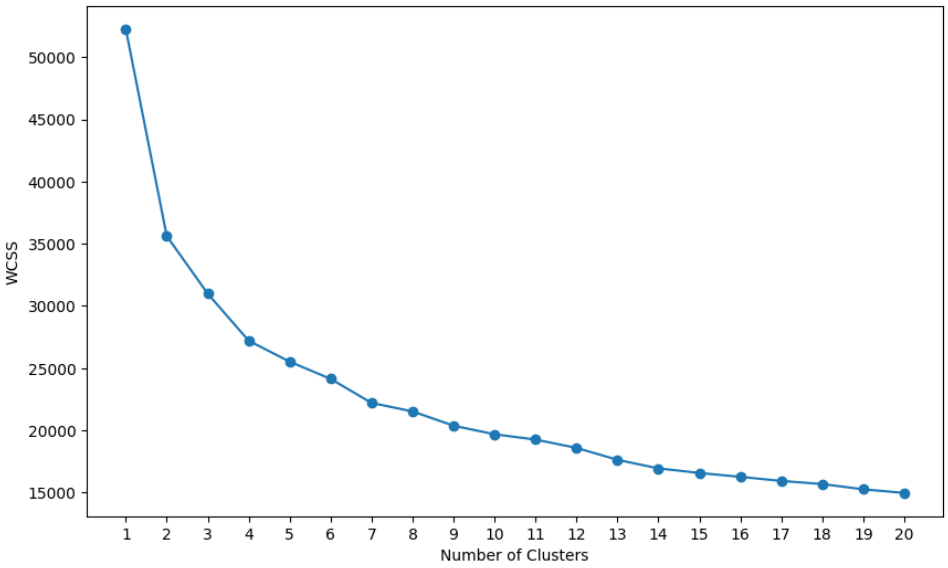

*Figure 4.8 - **Elbow Method for Optimal Number of Clusters**. Graph illustrating the within-cluster sum of squares (WCSS) as a function of the number of clusters, identifying the optimal cluster count by observing the "elbow point" where the rate of decrease slows significantly.*

4.2 Modelling

4.2.1 Modelling Results

This section show results of three machine learning models used for game outcome classification; Random Forest, Gradient Boosting and XGBoost. After all model preparation the model was trained using 162 features (including player type clusters) and 1520 matches (season 2-5). The model was tested on 380 matches (season 6) and validated on 380 matches (season 7).

Looking at Figure 4.1 (left) of target distribution, the dataset was quite balanced. Given this balance, accuracy became a more reliable and meaningful metric. Table 4.1 of global model metrics showed that Random Forest Achieved an Accuracy of 51,3% falling behind the gradient boosting models. The best performing model was the XGBoost model with a slightly higher accuracy over Gradient Boosting (55,2% vs 54,7%). For the other metrics in the table, the models perform quite similar over the board, with the exception of Random Forest having a 10% higher precision than the Gradient Boosting Models. For this prediction task, precision was less critical as it measured the quality of positive predictions, which would be more important if the cost of a false positive was costly, such as in medical predictions or fraud detection.

*Table 4.1 - **Model Performance Metrics.** Comparison of multiple performance metrics (Accuracy, Precision, Recall, F1, and ROC AUC) for Random Forest, Gradient Boosting, and XGBoost models, highlighting differences in their predictive capabilities.*

Model	Accuracy	Precision	Recall	F1	ROC AUC (Average)
Random Forest	51,3%	0,45	0,44	0,44	0,63
Gradient Boosting	54,7%	0,35	0,45	0,4	0,64
xGBoost	55,2%	0,35	0,46	0,4	0,64

4.2.1.1 ROC AUC

Investigating the Receiving Operator Curve (ROC AUC) of the models in Figure 4.9, revealed that the models perform similarly. All three models performed the best in distinguishing the classes for AwayWins, followed by HomeWins and Draws. The three of the models had issues predicting draws, with the Random Forest having a slight edge over the two other models in predicting draws, while the others performed better in predicting away- and homewins. Looking at the averages of the models, XGBoost and Gradient Boosting had a slightly higher ROC_AUC of 0,64 against 0,63 for Random Forest.

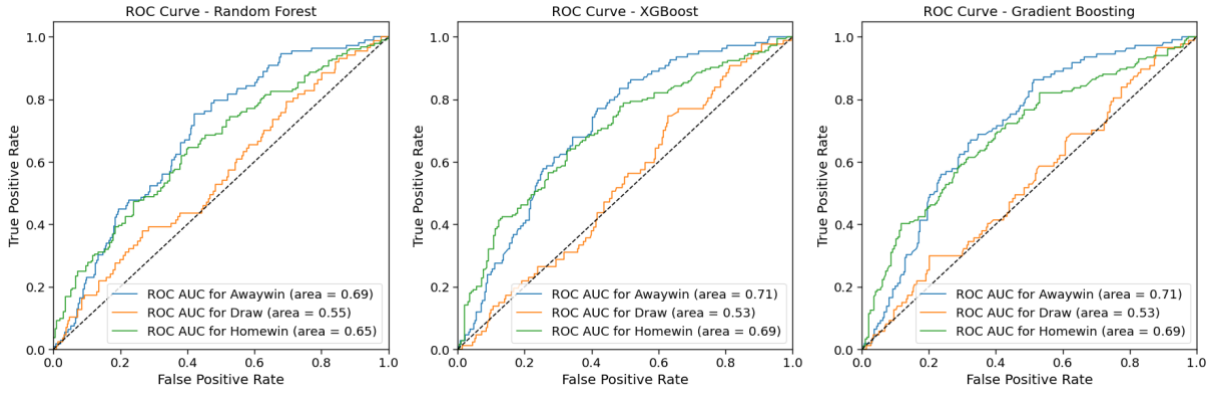

*Figure 4.9 - **Model ROC Curves for Home Win, Away Win, and Draw Predictions.** Comparison of ROC curves for Random Forest (*left*), XGBoost (*middle*), and Gradient Boosting (*right*) models, showing the true positive rate versus false positive rate for each match outcome (Home Win, Away Win, Draw) along with their respective AUC scores.*

4.2.1.2 Confusion Matrix

In Figure 4.10, the confusion matrixes for each model are plotted, providing insights into the models' strengths in predicting Home, Draw or AwayWins. All three models demonstrated strong capabilities of predicting HomeWins with Gradient Boosting predicting 150 correct, followed by XGBoost and Random Forest with 145 and 126. HomeWins was the outcome that was predicted correctly most times by all models. Following HomeWins, the models also

showed good capabilities in predicting AwayWins with XGBoost taking the lead on predicting 65 correct, followed by 58 and 57 by Gradient Boosting and Random Forest. Furthermore, interestingly the models struggled to predict draws, and the only model that was able to correctly predict draws was Random Forest. Looking at the ROC Curve in figure 4.9 (left), the values were close to a random guess with the area under the curve being close to 0,5, thus not being much better than the gradient boosting models.

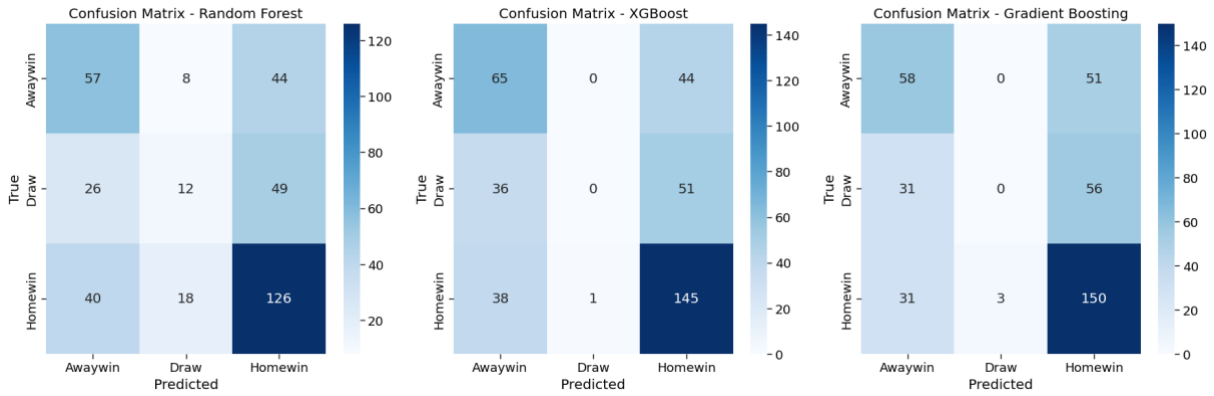

Figure 4.10 - ***Confusion Matrices on Test Set.*** *Visualization of confusion matrices for Random Forest (left), XGBoost (middle), and Gradient Boosting (right) models, showing the distribution of predicted versus actual outcomes (Home Win, Away Win, Draw) for the test dataset.*

4.2.2 Validation set

Model evaluation on unseen data (validation dataset) showed an accuracy of 58%. Which was in light with metrics reported for training data, with a slight increase for all outcomes. This was also reflected in the confusion matrix in Figure 4.11 (right), showing similar results, where HomeWins were most present and fewer AwayWins. Looking at Figure 4.11 (left) the ROC Curve was also similar to the ones in Figure 4.9 (right), showing that the model performs well on unseen data.

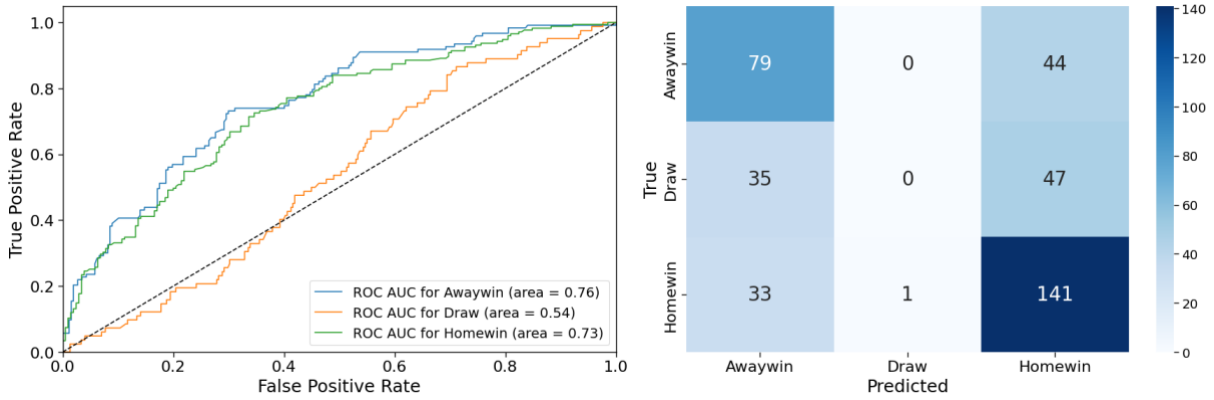

Figure 4.11 - **Validation Set Metrics for XGBoost**. (left): ROC AUC curve showing model performance for predicting Away Win, Draw, and Home Win. (right): Confusion matrix illustrating the distribution of true versus predicted outcomes on the validation set.

4.2.3 Betting Odds

As the dataset also had the betting odds of each outcome for all observations. These odds come from the betting provider Bet 365. After normalizing the odds, and calibrating the predictions into probabilities seen in figure 4.12 (left, right), that for Home- and AwayWins was dispersed along the red diagonal line and showed no significant outliers. For draws the predicted probabilities were generally lower but also showed no signs of significant outliers (Figure 4.12 middle). Furthermore, the Mean Average Error and Mean Squared Error was calculated, as seen in Table 4.2. Looking at the values it is seen that the calibrated probabilities were similar of the probabilities made by the odds.

Table 4.2 - ***Comparison of Mean Average Error (MAE) and Mean Squared Error (MSE) of calibrated model and the betting odds***

Category	MAE	MSE
H	0.0648	0.00682
D	0.0343	0.00172
A	0.0620	0.00619

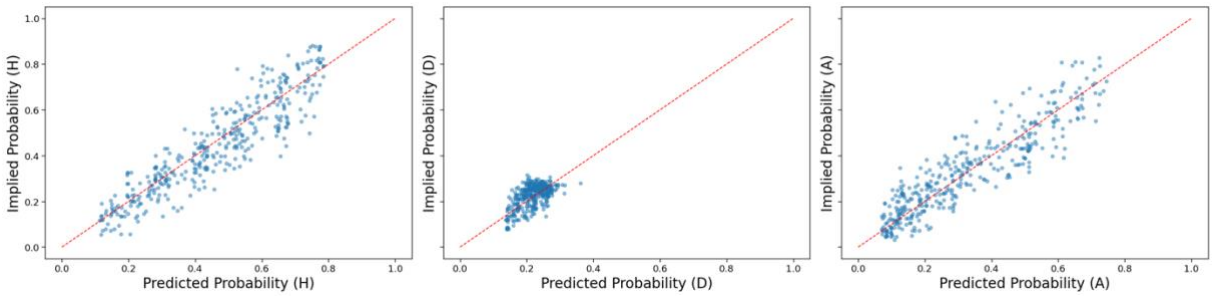

Figure 4.12 - ***Predicted vs Implied Probabilities.*** *Scatterplots comparing predicted probabilities with implied probabilities for left: Home Win, middle: Draw, and right: Away Win, highlighting calibration and alignment between model predictions and bookmaker odds.*

Finally, to compare the predictions of the odds and the model, a confusion matrix based off the bookmakers (Bet 365) probabilities was created. In figure 4.13 this confusion matrix showed that the bookmakers get a 60% accuracy, which is 2 % better than the XGBoost model. One notable finding here is that the odds also didn't classify any of the matches in the validation set as draws, due to the probabilities of home- and awaywins being higher than the probability of a draw.

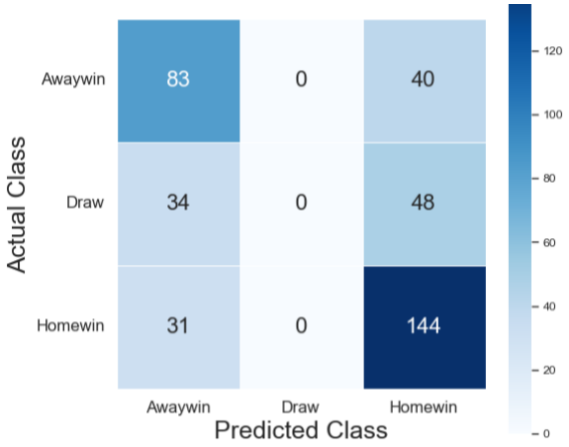

*Figure 4.13 - **Confusion Matrix Based on Bookmaker Odds.** Confusion matrix illustrating the predictions made by bookmaker probabilities (Bet365) compared to actual match outcomes on the validation set*

4.3 SHAP for global explanations

Shap analysis is a valuable tool to understand the factors that drive predictions into different classes. The SHAP summary plots for each class revealed which features contributed the most significantly to the model's predictions. The plots (Figure 4.12, 4.13 and 4.14) show the top 20 features for each class globally. Looking at feature importance across all categories, the plots show that the feature with highest predictive power was ELO Difference (elo_diff), contributing to both positive and negative outcomes. Additionally, average age of different positions affected outcomes, indicated that player experience in different roles, had different predictive powers. In the next subsections a detailed analysis for each one of the 3 outcomes (HomeWin, Draw and AwayWin) is provided.

4.3.1 Class A

For class A, the ELO Difference (elo_diff) had substantial impacts on the model's performance (Figure 4.14). A high value of elo_diff was assosciated with a negative prediction, while a lower value pushed the probability towards a negative prediction. Following that it was seen that the average age of offensive midfielders for the awayteam (away_Avg_Age_OM) and the average age of the starting 11 for the away team (away_Avg_Age_Starting_11) also had high impacts on the outcomes, although a higher average age of offensive midfielders pushed the prediction towards negative, and a lower starting 11 average age pushed the predictions towards a positive prediction.

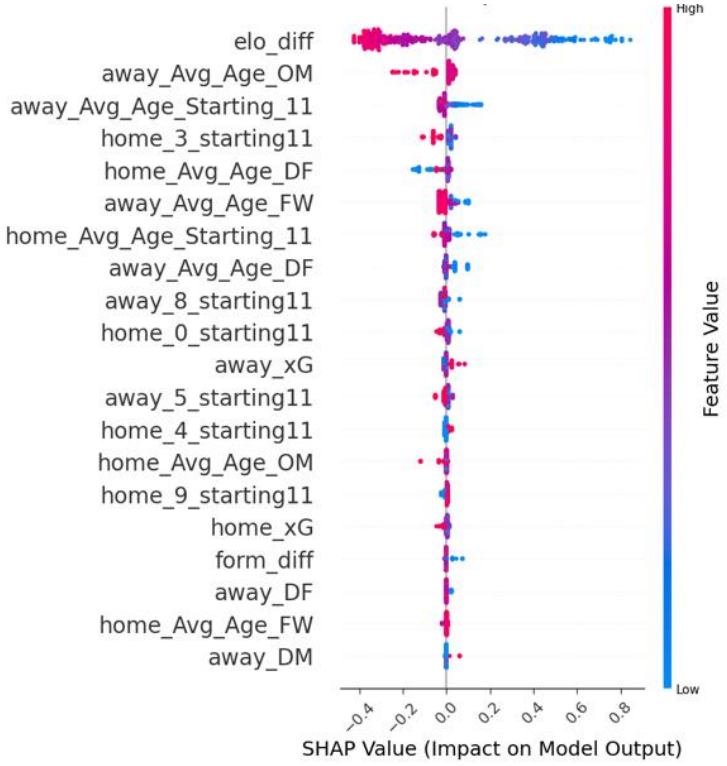

Figure 4.14 - **Global SHAP Summary Plot (Away Wins).** SHAP summary plot illustrating the impact of individual features on the model's predictions for Away Win outcomes, with feature importance ranked by mean absolute SHAP values and feature values represented along the color gradient.

4.3.2 Class D

The SHAP summary plot for Draws demonstrated a higher sensitivity to formation and defensive metrics (Figure 4.15). The features related to this class such as Home and Away defensive players, and formations suggested that in this class, the model considered that the tactical structure and lineup, mattered more than in the two other classes. It was also identified that ELO difference was an influential feature here.

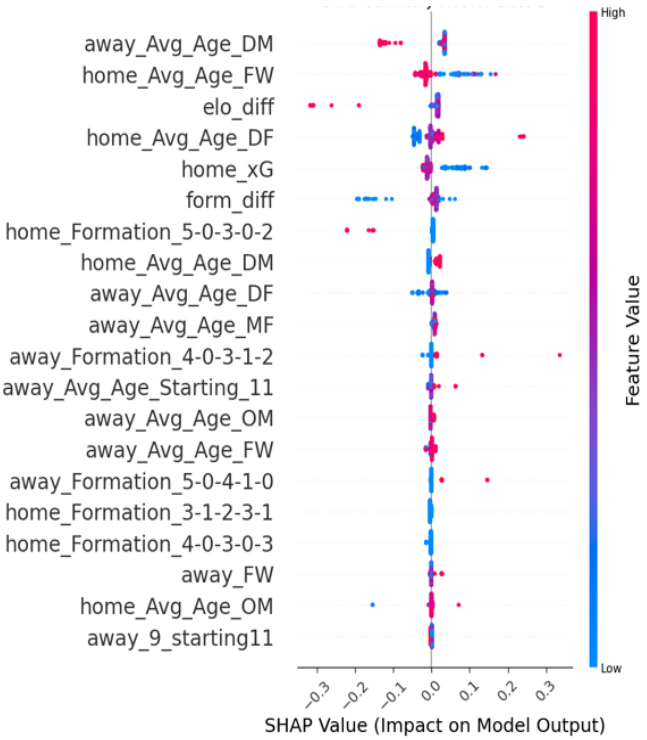

Figure 4.15 - **Global SHAP Summary Plot (Draws).** SHAP summary plot illustrating the impact of individual features on the model's predictions for Draw outcomes, with feature importance ranked by mean absolute SHAP values and feature values represented along the color gradient.

4.3.3 Class H

For the classification of HomeWins, the Expected Goals (xG) for the away and home teams stood out as some of the most influential features, after the ELO difference (Figure 4.16). A higher Away Expected Goals (away_xg) affected the model's outcome towards a negative prediction, while a higher Home Expected Goals (home_xg) pushed the model towards a HomeWin. For this outcome ELO Difference was opposite of what was seen for AwayWins. A higher elo_diff pushed the prediction towards a positive prediction, while a lower difference pulled more towards a negative prediction. Moreover, alongside these features, features such as average ages and player clustering were identified to have a high impact on the model outcome.

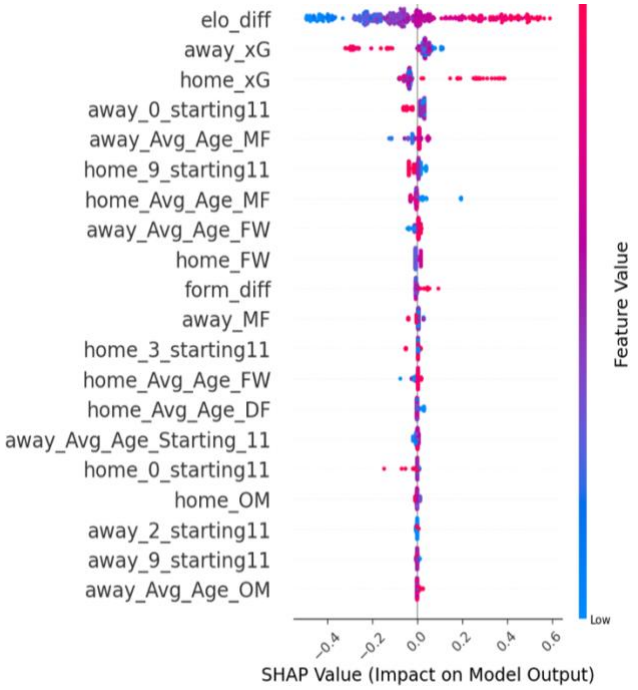

Figure 4.16 - **Global SHAP Summary Plot (Home Wins).** SHAP summary plot illustrating the impact of individual features on the model's predictions for Home Win outcomes, with feature importance ranked by mean absolute SHAP values and feature values represented along the color gradient.

4.4 SHAP for local explanations

In addition to global explanations, local examination was also conducted to understand what led to the outcome of specific predictions. By leveraging SHAP, it was possible to plot the SHAP values for each prediction. In figure 4.17 a SHAP waterfall plot for observation 55 was used as an example. The model predicted a HomeWin for this case and can be seen in Figure 7.1 (Appendix). When examining these plots, it is seen how the different features drew the prediction towards a negative or positive prediction for each class. For example the ELO Difference drew the prediction towards a negative prediction for Class A (Figure 4.15). For class D seen in Figure 7.2 (Appendix) the top 7 variables pushed more towards a positive prediction. For HomeWin the starting 11 of the away team drew towards a negative prediction and away xG pushed more for a positive prediction. The expected outcomes for the classes were ordered as 0,268 (AwayWin) , 0,379 (Draw) and 0,646 (HomeWin).

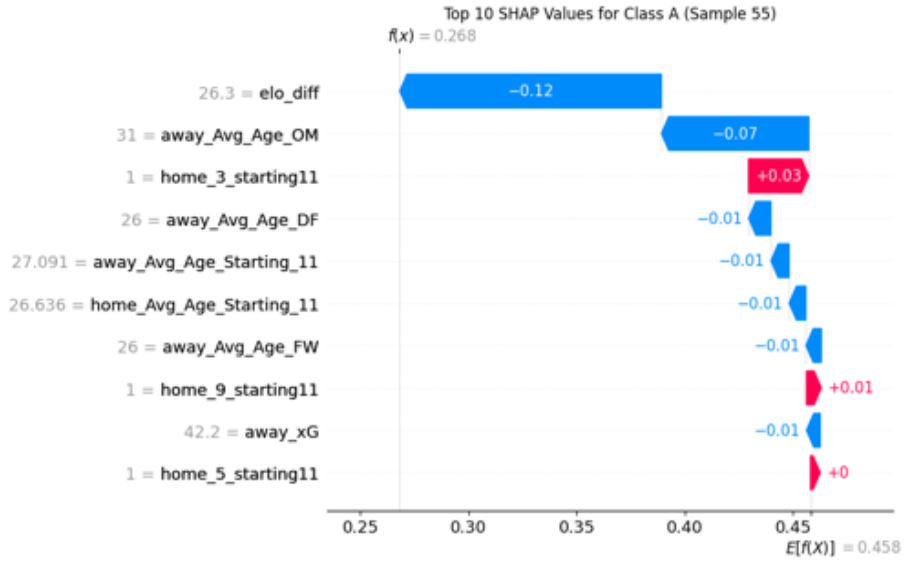

*Figure 4.17 - **SHAP Waterfall Plot for Away Win (Sample 55).** Waterfall plot showing the contribution of the top 10 features to the predicted probability of an Away Win for a specific instance (Sample 55). Positive and negative impacts on the prediction are illustrated, with the base value adjusted to the final prediction value.*

4.5 SHAP for specific teams

I addition to investigation of local and global explanations, custom insights and plots based on different features were created. As an example in figure 4.18 the top 10 features for a given team, in this case Manchester United, and the features that affect them to win, loose and draw at home. Away_xG, player-cluster 0 for away team and ELO difference affect the outcome of a HomeWin prediction. Loosing at home, ELO difference is the highest contributor and for draws, average ages and ELO difference affected the outcome of matches. These plots can be generated in multiple combinations with features, and for different teams to draw actionable insights

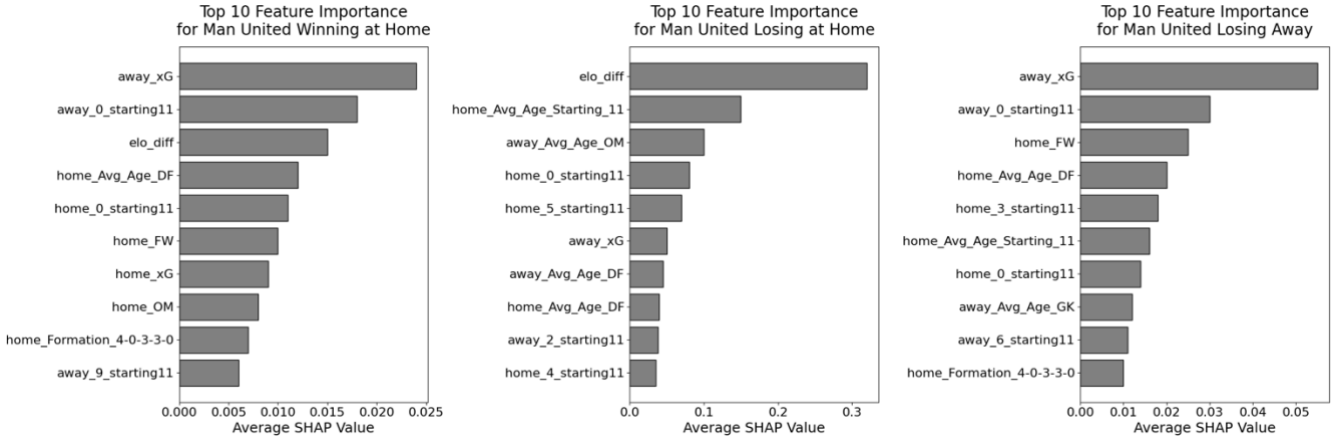

*Figure 4.18 - **Top 10 Feature Importance for Different Scenarios**. Left: Manchester United winning at home, Middle: Manchester United losing at home, Right: Manchester United losing away. Average SHAP values are used to rank features influencing match outcomes, highlighting the most impactful factors for each scenario.*

4.6 Formation simulations

By testing all different combinations, 1936 different combinations were tested for one specific prediction to provide an indication of what formations work best for one specific match. In Table 4.3, the combinations were ordered by the probability of an away win, and the top two observations were selected. Table 4.4 shows the bottom two observations. The tables were ordered by the probability of an away win. Through this process, it was observed that the probabilities changed significantly for some formations. For this specific match, the probability of an away win changed by 3.4%, 12% for draws, and 7.8% for a HomeWin. In this way, it was possible to identify which formation the model considered the best for any specific outcome and match.

*Table 4.3 - **Local Top 2 Formations and Outcome Probabilities for Away Win Predictions.** Home and away team formations ranked by the local probability of an Away Win (Prob A) for a specific match, alongside probabilities for a Draw (Prob D) and a Home Win (Prob H).*

Top 2 Ordered by Prob A				
Home Formation	**Away Formation**	**Prob A**	**Prob D**	**Prob H**
home_Formation_5-0-3-0-2	away_Formation_4-3-1-2-0	0.211736	0.219637	0.568627
home_Formation_5-0-3-0-2	away_Formation_3-2-2-0-3	0.203138	0.210718	0.586144

*Table 4.4 - **Local Bottom 2 Formations and Outcome Probabilities for Away Win Predictions.** Home and away team formations with the lowest local probabilities for an Away Win (Prob A), accompanied by probabilities for a Draw (Prob D) and a Home Win (Prob H).*

Bottom 2 Ordered by Prob A				
Home Formation	**Away Formation**	**Prob A**	**Prob D**	**Prob H**
home_Formation_5-0-2-3-0	away_Formation_5-0-4-1-0	0.177521	0.310249	0.512230
home_Formation_3-1-2-3-1	away_Formation_5-0-4-1-0	0.170063	0.339226	0.490710

To find an overall best formation, the predictions were ran for 100 samples resulting in 193600 predictions and calculated the average scores the model gave each outcome. To be consistent with the tables, table 4.5 and 4.6 are ordered by Prob A, but could of course be ordered in any way desired. It was identified that the best formation for away wins overall was away 4-3-1-2-0, and the worst away formation, is 5-0-4-1-0 according to model scores. More filtering and extensive tables would also allow us to investigate specific formations and outcomes, both locally and globally. This analysis allowed us to get a better understanding of overall formation performance but also match specific formation performance.

*Table 4.5 - **Global Top 2 Formations and Outcome Probabilities for Away Win Predictions.** Home and away team formations ranked by the global probability of an Away Win (Prob A) for a specific match, alongside probabilities for a Draw (Prob D) and a Home Win (Prob H).*

Top 2 Ordered by Prob A				
Home Formation	**Away Formation**	**Prob A**	**Prob D**	**Prob H**
home_Formation_5-0-3-0-2	away_Formation_4-3-1-2-0	0.358390	0.207186	0.434424
home_Formation_5-0-3-0-2	away_Formation_2-1-4-2-1	0.358347	0.207080	0.434573

*Table 4.6 - **Global Bottom 2 Formations and Outcome Probabilities for Away Win Predictions.** Home and away team formations with the lowest global probabilities for an Away Win (Prob A), accompanied by probabilities for a Draw (Prob D) and a Home Win (Prob H).*

Bottom 2 Ordered by Prob A				
Home Formation	**Away Formation**	**Prob A**	**Prob D**	**Prob H**
home_Formation_5-0-2-3-0	away_Formation_5-0-4-1-0	0.331630	0.261130	0.407239
home_Formation_3-1-2-3-1	away_Formation_5-0-4-1-0	0.327780	0.269250	0.402970

5 Discussion

This study sought to address the following research questions: (i) **How effectively can machine learning models predict English Premier League match outcomes using historic match, team, and player-specific data**, (ii) **How can explainable AI be used to provide actionable insights for managerial decision-making based on player positional data?**

5.1 Summary of Key Findings

This study resulted in three major findings in line with the first research question. These findings include: (i) the best performing model in comparison to previous literature, (ii) the odds performance against the XGBoost model, and (iii) the feature importance derived from SHAP values. The most comparable study for performance was Baboota & Kaur's study (Baboota & Kaur, 2019). Baboota & Kaur used a Gradient Boosting model reaching 59% accuracy predicting in the English Premier league while also accounting for draws. This is in line with the work presented here, with an accuracy of 55% on test set and 58% on the validation set. Furthermore, looking at both confusion matrixes in figure 4.10 middle and figure 4.13 the predicted matches are similar of the ones from the odds, where this study comes slightly behind on predicting HomeWins and AwayWins. Both models struggle to predict draws. Lastly, it was identified that match, player and team specific data affect the outcome of the model's conclusion in the SHAP plots 4.14, 4.15 and 4.16.

In relation to the second research question this study has identified two major findings that align with the research question. These findings were how SHAP allowed us to understand underlying factors of the model's conclusion, and how formation types affect the probability of outcomes. The SHAP plots allowed for both global and local understanding of what factors led to the predicted outcomes, and identified the most prominent variables as ELO, both locally and globally. Team and player statistics also had a significant impact on the outcomes of matches, such as expected goals, age and form.

5.2 Data Insights

The EDA of this study revealed insights into the different datasets used in this study. The match dataset showed little target imbalance seen in figure 4.1 left. This target balance did not require target balancing techniques due to the small difference in outcomes. Figure 4.2 right showed

the average implied probability of outcomes against actual outcomes, and showed a close relationship, laying grounds for using the odds as a further comparison measure. The match data showed that over time the distribution of results was consistent, with the exception of season 4 where there were slightly more AwayWins, than HomeWins.

The Match and ELO datasets had time series data that had to be considered to avoid potential data leaks. This was dealt with by making sure no updates were done to the features before the match occurred, eliminating the risk of data leaks. The player statistics data was static for each season, and therefore different from the other datasets. This led to having to adjust the data to be able to enrich the dataset used for modelling. This was done by aggregating team level metrics based on individual player performances. The correlation matrix in figure 4.5 (left) revealed strong correlation between goal metrics and progressive metrics. Figure 4.5 (right) dove further into these insights and revealed that different player positions had vastly different distributions of progressive features. This layed the ground for k-means clustering to draw more information about the different player types and allowed us to classify them to clusters, illustrated in figure 4.7. This approach showed some clusters centered together towards the middle, and suggest similar characteristics between these players, while some of the clusters were more dispersed suggesting greater diversity in player attributes. In line with the first research question, the number of goals scored is what effectively affects a match outcome, and therefore these statistics suggested that some of these metrics may aid in classifying the outcome.

5.3 Modelling Results

Accuracy was chosen as the main metric of performance, justified by the target balance. XGBoost performed the best out of the three models developed. The model reached 55% accuracy on the test set and 58% on the validation set. This was comparable to the study of Baboota & Kaur reaching 59% accuracy on a similar prediction task in the English Premier League using a Gradient Boosting Model (Baboota & Kaur, 2019). The differences here could be explained by them having a better model, or a season where the outcomes were easier to predict. Similarly to their study, the models struggled in predicting draws, with both gradient boosting models classifying zero draws in the test set. Meanwhile the Random Forest Model managed to classify 12 draws correctly, but looking at figure 4.7 (left), the predicted draws are close to random guessing. This is also seen in the study of Baboota & Kaur across all the tested

models. The results were compared with the odds (table 4.2, and figure 4.12), revealed that predictions differed between 6 and 4% from the odds on average which is a positive indication of how the model performs. This alignment demonstrates that the model captures trends in the data that are consistent with real-world probabilities represented by the odds. Moreover, after converting the odds to a classification model, it achieved 60% accuracy on the validation set, with no predicted draws which suggests that there is always a higher probability of a HomeWin or AwayWin. This supports both the study of Baboota & Kaur as well as this study.

Other studies have achieved higher accuracies within the field of football match prediction, such as a Bayesian Network by Joseph et al, and a Logistic regression by Prasetio and Harlili gaining 59,21% and 69,5% accuracy (Joseph et al., 2006; Prasetio & Harlili, 2016). Although reaching higher accuracies, the studies had limitations such as generalization to multiple teams or predicting only a subset of outcomes, that can lead to higher accuracies. Josepth et al, used a Bayesian Network where predictions were specific for one team, and the Logistic Regression by Prasetio and Harlili did not account for draws. If the task was to only predict home and awaywins the model in this study would have an accuracy of 73,8% in the validation set.

5.4 SHAP Analysis

SHAP was applied in this study to understand the conclusions of our machine learning model. A significant finding here was that ELO Difference (elo_diff) was the most influential feature in predicting Home and AwayWins, which aligned with the study of Hvattum and Arntzen which used ELO to predict the outcome of football matches in trying to beat the odds (Hvattum & Arntzen, 2010). This underlines the power of using ELO as a feature for predicting match outcomes, as it uses both current form and historic form. In addition to ELO it was found that variables like player clusters, expected goals and age had a high impact on outcome, which supports the first research question as these features came from multiple sources of data (Player-, Match-, ELO- data). SHAP was also applied locally to illustrate the conclusion of the model for specific outcomes. This was done to support the second research question on how one may draw managerial decision-making material from this. Figure 4.17 illustrates how the model reasoned and moved the probability towards a positive or negative classification. Using this prediction, a manager would be able to see what features drew the match toward a loss or a win. For this instance, the age of the away team's offensive midfielders, drew the predictions towards a negative classification, and the manager could then apply changes to the lineup if younger offensive midfielders were in the stall. Other features such as player clusters, formation and

lineups could also be altered before the match by using SHAP values to enhance decision-making.

To better understand of some of these SHAP values Figure 4.14 show a global importance summary plot. The average age of away offensive midfielders was the second most influental feature for awaywins. A study conducted on when football players peak was conducted to study when different positions peak (Dendir, 2016). The study highlights that forwards peak around 25, midfielders 25-27 and defenders at 27, where defenders maintain their performance the longest. The decline in performance after their peak years may explain why we see the influence of age on the outcome of our model's prediction where a higher age for certain positions such as away_avg_age_OM show that a higher age pulls the prediction towards a negative classification while an overall lower average age on the away team pulls towards a positive prediction.

SHAP values were also aggregated on team level (figure 4.18) showing the top 10 feature importance for specific scenarios. These plots were inspired by a study on team level performance in the premier league using leveraging SHAP (Moustakidis et al., 2023). This way a manager could identify weaknesses and strengths with their own team as well as opposing teams to be able to identify strategies to have a better chance at succeeding in matches against the teams. While the feature importance values in these plots appear relatively low, their interpretation remains valid and useful because the model demonstrates performance consistent with prior studies and aligns closely with bookmaker predictions.

5.5 Data Simulations

From the player statistics dataset, formations were created for each team. By testing all different combinations of formations, it was possible to identify which formations had the highest probability of a given outcome locally and globally. It was identified that the percentages changed significantly when changing formations. In the example, the probability of an away win changed by 3.4%, 12% for draws, and 7.8% for home wins. This indicates that a manager could have increased their probability of winning by 7.8% (if using the worst-ranked formation) simply by changing formations. In line with the research question, this demonstrates how the results of this study could be applied in a production setting where the best formations for specific matches can be predicted.

5.6 Limitations

This chapter outlines the key limitations of the present study, acknowledging where it's scope, methodology and data may have constrained the robustness and generalizability of the findings. While the analysis and models have provided valuable insights there have been limitations that limit the robustness of this study.

One of the key limitations of this study was the Player Statistics Dataset. As the dataset only provided a picture of statistics for the whole season, it had to be lagged by one season, before joining any of the created features to the other datasets. Because of this, a lot of noise was not captured by the model and potentially decreasing the model's accuracy and interpretability. It also influenced the quality of the variables created from the dataset, and how they can be used in the interpretation of SHAP values for decision making as features, such as formation change throughout the season. Furthermore, another limitation of this study it only uses results as match specific features and does not consider in-game features such as yellow- and red cards, tackles, and shots on target. These variables could contain valuable information for predicting. Another limitation of this study was that it does not consider any other factors than what happens in the league, such as travel fatigue, travel distance, international breaks, injuries and transfers which have been used in previous studies (Prasetio & Harlili, 2016).

This study was based on the English Premier League, which introduced a limitation of usage of the model. A study conducted on the differences between the top 5 football leagues revealed that there are different playstyles in the leagues (Yi et al., 2019). This indicated that the model may perform worse if applied to another football league than the English Premier League. In addition to this, the study is based on male football players and may not be representative of women's football. A study comparing home advantage in men's and women's football in Europe concluded that home advantage exists for women's football, but at a higher degree for men's football (Pollard & Gómez, 2014). This raises the same issue mentioned above, and the model could overestimate the importance of home advantage if applied to women's football.

Lastly, a limitation that is also seen in other studies is the ability to predict draws (Baboota & Kaur, 2019). As seen in figure 4.1 left, 22,6% of the matches result in a draw. Predicting draws was a difficult task, and the odds also classified the probability of a draw always being lower than a HomeWin or AwayWin. This limits the achievable accuracy with the current machine learning model.

5.7 Conclusion

In conclusion, the XGBoost model provided valuable insights and good accuracy in classifying Premier League Matches as HomeWins or AwayWins. The model had a limitation in classifying matches as draws, but previous literature reveal that the model is performing in line with what has been researched before (Baboota & Kaur, 2019; Prasetio & Harlili, 2016). The inclusion of ELO data, player and team level data provided features that influenced the model's prediction according to our SHAP plots and is in line with the research question. On how effectively one can predict, the model had 55% (test) and 58% (validation) accuracy with the current features, and it is seen that match, player and team specific features all affect the outcome of our model.

SHAP values allowed for understanding of the conclusions of the model, and as mentioned in section 5.4 it is possible to draw insightful decisions based on this data, where a team manager could both use global explanations to better understand overall predictions, but also local predictions where a team may be micromanaged with the altering of formations and players. Lastly, data simulations also provided both local and global insight into what formations the model deem as the best models. Notably local interpretations of formations allow for changing formations to increase the probability of a desired outcome.

Concluding, the current methodology allowed for efficient predictions of the English Premier League when predicting away-, and homewins, but the model struggled in predicting draws. The inclusion of team and player statistics provided insightful information and affects the model's conclusion based on SHAP values. Drawing insightful decisions from a managerial concept is possible, but should be considered with caution, especially due to the noise present in the player data. To ensure meaningful actionable insights player level data should be gathered on a higher granularity.

5.8 Future Work

This study has laid a solid foundation in the field of predicting football matches in the English Premier League, but there are several areas that warrant further exploration. Key limitations such as player data being static, provide opportunities to refine and enhance the methodologies applied. The research can build on the findings of how formations, team and player level data

in more granular format, can enhance the performance of the model, but also provide a better explainability of the model.

An interesting approach for further study would be to turn the model into a binary classification model. Due to the nature of the research's goal being to draw managerial decision-making, predicting a draw is not critical, while enhancing the chances of a win being important for stakeholders within a football club. Such as in the study of Prasetio and Harlili using a Logistic Regression and achieving 69,21 % accuracy (Prasetio & Harlili, 2016). It was seen that this study would have reached a 73,8% accuracy if the model was a binary classification model (predicting draws, and wins). Combining knowledge from this study with a binary classification method, could also enhance the ability to interpret the results. Deep Learning approaches, such as Neural Networks have been mentioned in the literature review and have been used to predict the number of goals in football matches with promising results (Arabzad et al., 2014). There is not much research applying LSTM to football match predictions, but this type of Neural Network has been proven to perform well on time series data, and it could an interesting field of research combining the knowledge obtained from this study (Siami-Namini et al., 2018). Combining the usage of LSTM, SHAP, Granular Player Data and Binary Classification could be a future work.

6 Bibliography

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Issue 8, pp. 1–12). MDPI AG. https://doi.org/10.3390/electronics9081295

Arabzad, S. M., Araghi, M. E. T., Sadi-Nezhad, S., & Ghofrani, N. (2014). Football Match Results Prediction Using Artificial Neural Networks; The Case of Iran Pro League. In *Journal of Applied Research on Industrial Engineering* (Vol. 1, Issue 3).

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Böken, B. (2021). On the appropriateness of Platt scaling in classifier calibration. *Information Systems*, *95*. https://doi.org/10.1016/j.is.2020.101641

Breiman, L. (2001). *Random Forests* (Vol. 45).

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Dendir, S. (2016). When do soccer players peak? A note. *Journal of Sports Analytics*, *2*(2), 89–105. https://doi.org/10.3233/jsa-160021

Ding, C., & He, X. (2004). *K-means Clustering via Principal Component Analysis*.

Dixon, M. J., & Coles, S. G. (1997). *Modelling Association Football Scores and Inefficiencies in the Football Betting Market*.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3561048

Edwards, L., Veale, M., Welbl, J., Kleek, M. Van, Binns, R., Lane, G., & Henderson, T. (2017). *SLAVE TO THE ALGORITHM? WHY A "RIGHT TO AN EXPLANATION" IS PROBABLY NOT THE REMEDY YOU ARE LOOKING FOR* (Vol. 16). https://perma.cc/PJX2-XT7X];

Elo, A. E. (1978). *The Rating of Chess Players, Past and Present*.

FBREF. (n.d.). *FBREF*. Https://Fbref.Com/En/Comps/9/Stats/Premier-League-Stats.

Football-Data. (n.d.). *football-data.co.uk*. Https://Football-Data.Co.Uk/Englandm.Php.

Grimes, D. A., & Schulz, K. F. (2008). Making Sense of Odds and Odds Ratios. In *Obstet Gynecol* (Vol. 111).

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460–470. https://doi.org/10.1016/j.ijforecast.2009.10.002

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553. https://doi.org/10.1016/j.knosys.2006.04.011

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6). www.ijarcsms.com

Lars Schiefler. (n.d.). *Football Club Elo Ratings*. Http://Clubelo.Com/.

Lundberg, S., Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/arXiv.1705.07874

Maher, M. J. (1982). *Modelling association football scores*.

Moustakidis, S., Plakias, S., Kokkotis, C., Tsatalas, T., & Tsaopoulos, D. (2023). Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. *Future Internet*, *15*(5). https://doi.org/10.3390/fi15050174

Pollard, R., & Gómez, M. A. (2014). Comparison of home advantage in men's and women's football leagues in Europe. *European Journal of Sport Science*, *14*(SUPPL.1). https://doi.org/10.1080/17461391.2011.651490

Prasetio, D., & Harlili, D. (2016). *Predicting Football Match Results with Logistic Regression*. IEEE.

premierleague.com. (n.d.). *Premier League Stats*. Https://Www.Premierleague.Com/Stats/Clarification.

Provost, F., & Fawcett, T. (2013). *Data Science for Business*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning*. http://arxiv.org/abs/1606.05386

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Spann, M., & Skiera, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55–72. https://doi.org/10.1002/for.1091

Statsperform. (n.d.). *Opta*. Https://Www.Statsperform.Com/.

Sun, D. (2023). An Overview of Machine Learning Applications in the Football Field. *Applied and Computational Engineering*, *8*(1), 318–322. https://doi.org/10.54254/2755-2721/8/20230178

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wang, Y., Liu, W., & Liu, X. (2022). Explainable AI techniques with application to NBA gameplay prediction. *Neurocomputing*, *483*, 59–71. https://doi.org/10.1016/j.neucom.2022.01.098

Yi, Q., Groom, R., Dai, C., Liu, H., & Gómez Ruano, M. Á. (2019). Differences in Technical Performance of Players From 'The Big Five' European Football Leagues in the UEFA Champions League. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02738

7 Appendix

*Table 7.1 - **Match Features.** Overview and description of all features in data from match specific features*

Date	The date on which the match was played.
Hometeam	The name of the team playing at home.
Awayteam	The name of the team playing away.
Fthg	Number of goals scored by the home team at full time.
Ftag	Number of goals scored by the away team at full time.
Ftr	The result of the match at full time: 'H' = Home Win, 'D' = Draw, 'A' = Away Win. **Target Variable**
Hthg	Number of goals scored by the home team at half time.
Htag	Number of goals scored by the away team at half time.
Htr	The result of the match at half time: 'H' = Home Lead, 'D' = Draw, 'A' = Away Lead.
B365h	Odds offered by Bet365 for a home win.
B365d	Odds offered by Bet365 for a draw.
B365a	Odds offered by Bet365 for an away win.
Season	The football season or year during which the data was collected.

*Table 7.2 - **Player Features.** Overview and description of all features in data from match specific features*

Player name	The full name of the player.
Nationality	The country the player represents or is eligible to play for internationally.
Pos	

	- The player's primary position(s) on the field (e.g., FW for Forward, MF for Midfielder, DF for Defender).
Squad	The club or team the player is affiliated with during the season.
Age	Players age at the start of the season
Born	The year the player was born.
MP	Total number of matches in which the player appeared.
Starts	Number of matches the player was in the starting lineup.
Min	Total minutes the player was on the pitch during the season.
90s	Number of full 90-minute matches the player has played (Minutes Played divided by 90).
Gls	Total number of goals scored by the player, including penalty kicks.
Ast	Total number of assists credited to the player.
G+A	Sum of goals and assists; a measure of total goal contributions.
G-PK	Total goals scored excluding penalty kicks.
PK	Number of penalty kicks successfully converted into goals.
PKatt	Total number of penalty kick attempts by the player.
CrdY	Number of yellow cards received by the player.
CrdR	Number of red cards received by the player.

xG	A metric that estimates the likelihood of a shot resulting in a goal based on factors like shot location and type.
npxG	Expected goals excluding penalty kicks; measures goal-scoring opportunities from open play and non-penalty set pieces.
xAG	The likelihood that a given pass will become a goal assist, based on factors like pass type and location.
npxG + xAG	Sum of npxG and xAG; evaluates a player's overall offensive contribution excluding penalties.
PrgC	Number of times the player carries the ball towards the opponent's goal by a significant distance (usually 5+ yards).
PrgP	Number of times the player moves the ball towards the opponent's goal through runs or dribbles.
Gls.1	Average number of goals scored by the player per 90 minutes played.
Ast.1	Average number of assists made by the player per 90 minutes played.
G+A.1	Combined goals and assists averaged per 90 minutes played.
G+A-PK	Combined non-penalty goals and assists averaged per 90 minutes played.
xG.1	Average expected goals per 90 minutes played.
xAG.1	Average expected assists per 90 minutes played.
xG+xAG	Combined expected goals and expected assists averaged per 90 minutes played.

npxG.1	Average non-penalty expected goals per 90 minutes played.
npxG+xAG.1	Sum of npxG.1 and xAG.1 per 90 minutes; measures offensive contributions excluding PKs.
Matches	Details or list of matches the player participated in during the season.
Season	The football season or year during which the data was collected.

*Table 7.3 - **ELO Features.** Overview and description of features from ELO data acquired from API*

Rank	The position of the club in the ranking based on their Elo rating or other performance metrics.
Club	The name of the football (soccer) club.
Country	The country where the club is based or the league in which it competes.
Level	The tier or division in the football league system where the club plays (e.g., 1 for top division).
Elo	The club's Elo rating, a measure of its strength based on historical match results.
From	The start date or period from which the Elo rating or ranking applies.
To	The end date or period until which the Elo rating or ranking applies.

*Table 7.4 - **Engineered Features** – Overview and description of all created features from Feature Engineering*

homewinstreak	3 Game winstreak
awaywinstreak	3 Game winstreak
home_season	Season data specific to the home team.

home_DF	Number of defenders in the home team's starting lineup.
home_DM	Number of defensive midfielders in the home team's starting lineup.
home_FW	Number of forwards in the home team's starting lineup.
home_GK	Number of goalkeepers in the home team's starting lineup
home_MF	Number of midfielders in the home team's starting lineup.
home_OM	Number of offensive midfielders in the home team's starting lineup.
home_xG	Expected goals for the home team in the match.
home_DF_total	Total defenders in the home team's squad.
home_DM_total	Total defensive midfielders in the home team's squad.
home_FW_total	Total forwards in the home team's squad.
home_GK_total	Total goalkeepers in the home team's squad.
home_MF_total	Total midfielders in the home team's squad.
home_OM_total	Total offensive midfielders in the home team's squad.
home_Avg_Age_Total_Squad	Average age of the entire home team squad.
home_Avg_Age_Starting_11	Average age of the home team's starting eleven.
home_Avg_Age_DF	Average age of defenders in the home team's starting eleven.
home_Avg_Age_DM	Average age of defensive midfielders in the home team's starting eleven.
home_Avg_Age_FW	Average age of forwards in the home team's starting eleven.
home_Avg_Age_GK	Average age of goalkeepers in the home team's starting eleven.

home_Avg_Age_MF	Average age of midfielders in the home team's starting eleven.
home_Avg_Age_OM	Average age of offensive midfielders in the home team's starting eleven.
home_Formation	Specific formation used by the home team in the match (e.g., 4-4-2).
home_Formation_Type	General type of formation (e.g., defensive, offensive, balanced).
home_0_starting11	Number of players in starting 11 in cluster
home_1_starting11	Number of players in starting 11 in cluster
home_2_starting11	Number of players in starting 11 in cluster
home_3_starting11	Number of players in starting 11 in cluster
home_4_starting11	Number of players in starting 11 in cluster
home_5_starting11	Number of players in starting 11 in cluster
home_6_starting11	Number of players in starting 11 in cluster
home_7_starting11	Number of players in starting 11 in cluster
home_8_starting11	Number of players in starting 11 in cluster
home_9_starting11	Number of players in starting 11 in cluster
home_10_starting11	Number of players in starting 11 in cluster
home_0_total	Number of players in cluster in squad
home_1_total	Number of players in cluster in squad
home_2_total	Number of players in cluster in squad
home_3_total	Number of players in cluster in squad
home_4_total	Number of players in cluster in squad
home_5_total	Number of players in cluster in squad
home_6_total	Number of players in cluster in squad
home_7_total	Number of players in cluster in squad
home_8_total	Number of players in cluster in squad
home_9_total	Number of players in cluster in squad
home_10_total	Number of players in cluster in squad
away_DF	Number of defenders in the away team's starting lineup.

away_DM	Number of defensive midfielders in the away team's starting lineup.
away_FW	Number of forwards in the away team's starting lineup.
away_GK	Number of goalkeepers in the away team's starting lineup (should always be 1).
away_MF	Number of midfielders in the away team's starting lineup.
away_OM	Number of offensive midfielders in the away team's starting lineup.
away_xG	Expected goals for the away team in the match.
away_DF_total	Total defenders in the away team's squad.
away_DM_total	Total defensive midfielders in the away team's squad.
away_FW_total	Total forwards in the away team's squad.
away_GK_total	Total goalkeepers in the away team's squad.
away_MF_total	Total midfielders in the away team's squad.
away_OM_total	Total offensive midfielders in the away team's squad.
away_Avg_Age_Total_Squad	Average age of the entire away team squad.
away_Avg_Age_Starting_11	Average age of the away team's starting eleven.
away_Avg_Age_DF	Average age of defenders in the away team's starting eleven.
away_Avg_Age_DM	Average age of defensive midfielders in the away team's starting eleven.
away_Avg_Age_FW	Average age of forwards in the away team's starting eleven.
away_Avg_Age_GK	Average age of goalkeepers in the away team's starting eleven.
away_Avg_Age_MF	Average age of midfielders in the away team's starting eleven.

away_Avg_Age_OM	Average age of offensive midfielders in the away team's starting eleven.
away_Formation	Specific formation used by the away team in the match (e.g., 4-4-2).
away_Formation_Type	General type of formation (e.g., defensive, offensive, balanced).
away_0_starting11	Number of players in starting 11 in cluster
away_1_starting11	Number of players in starting 11 in cluster
away_2_starting11	Number of players in starting 11 in cluster
away_3_starting11	Number of players in starting 11 in cluster
away_4_starting11	Number of players in starting 11 in cluster
away_5_starting11	Number of players in starting 11 in cluster
away_6_starting11	Number of players in starting 11 in cluster
away_7_starting11	Number of players in starting 11 in cluster
away_8_starting11	Number of players in starting 11 in cluster
away_9_starting11	Number of players in starting 11 in cluster
away_10_starting11	Number of players in starting 11 in cluster
away_0_total	Number of players in cluster in squad
away_1_total	Number of players in cluster in squad
away_2_total	Number of players in cluster in squad
away_3_total	Number of players in cluster in squad
away_4_total	Number of players in cluster in squad
away_5_total	Number of players in cluster in squad
away_6_total	Number of players in cluster in squad
away_7_total	Number of players in cluster in squad
away_8_total	Number of players in cluster in squad
away_9_total	Number of players in cluster in squad
away_10_total	Number of players in cluster in squad
home_missing_data	Proportion of missing data for the home team.
away_missing_data	Proportion of missing data for the away team.
form_diff	Difference in recent form between the home and away teams. Wins vs Losses

home_elo	ELO rating of hometeam
away_elo	ELO rating of awayteam
elo_diff	Difference in Elo ratings between the home and away teams.

*Table 7.5 - **k-Means Features.** Overview of all features used in the creation of clusters using k-Means Clustering*

Feature
Age
Min
90s
Gls
Ast
GA
GPK
PK
PKatt
CrdY
CrdR
xG
npxG
xAG
npxGxAG
PrgC
PrgP
PrgR
Gls1
Ast1
GA1
GPK1
GAPK
xG1

xAG1
xGxAG
npxG1
npxGxAG1

*Table 7.6 - **Modeling grid search overview.** Overview of grids searched for all hyperparameters for the models Random Forest, Gradient Boosting and XGBoost.*

Model	Hyperparameter	Values
Random Forest	N Estimators	50, 100, 200
	Max Depth	None, 10, 20, 30
	Min samples Split	2, 5, 10
Gradient Boosting	N Estimators	50, 100, 200
	Learning Rate	0.01, 0.1, 0.2
	Max Depth	3, 5, 7
XGBoost	N Estimators	50, 100, 200
	Learning Rate	0.01, 0.1, 0.2
	Max Depth	3, 5, 7

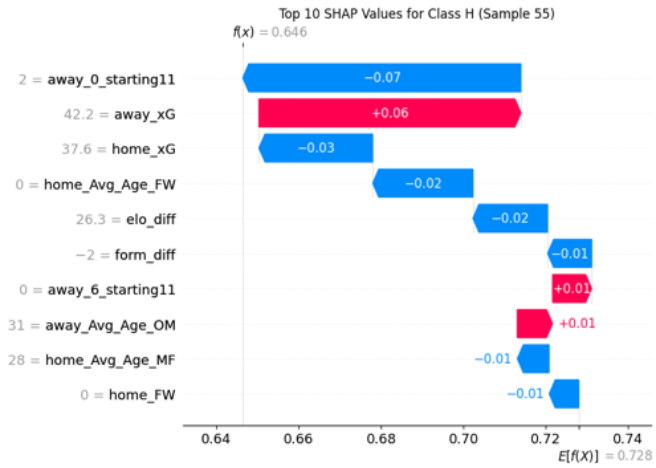

*Figure 7.1 - **SHAP Waterfall Plot for Home Win (Sample 55).** Waterfall plot showing the contribution of the top 10 features to the predicted probability of a Home Win for a specific instance (Sample 55). Positive and negative impacts on the prediction are illustrated, with the base value adjusted to the final prediction value.*

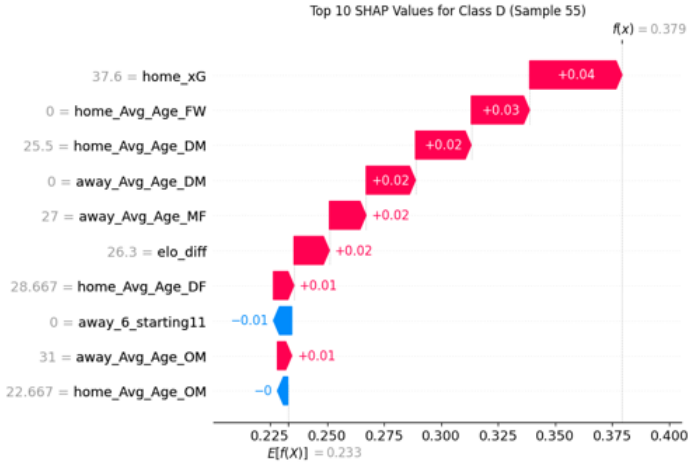

Figure 7.2 - ***SHAP Waterfall Plot for Draws (Sample 55).*** *Waterfall plot showing the contribution of the top 10 features to the predicted probability of a draw for a specific instance (Sample 55). Positive and negative impacts on the prediction are illustrated, with the base value adjusted to the final prediction value.*