National College of Ireland

Ball Possession metrics-oriented Analysis to Predict Football League Rankings

MSc Research Project
Data Analytics

Jaseem Jamal Panikkaveetil
Student ID: 21236828

School of Computing
National College of Ireland

Supervisor: Sasirekha Palaniswamy

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Jaseem Jamal Panikkaveetil
Student ID:	21236828
Programme:	Msc in Data Analytics
Year:	2023
Module:	Msc Research Project
Supervisor:	Sasirekha Palaniswamy
Submission Due Date:	14/12/2023
Project Title:	Ball Possession metrics-oriented Analysis to Predict Football League Rankings
Word Count:	8023
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jaseem Jamal Panikkaveetil
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Ball Possession metrics-oriented Analysis to Predict Football League Rankings

Jaseem Jamal Panikkaveetil
21236828

Abstract

The importance of ball possession from a strategic standpoint has drawn attention from coaches, analysts, and fans in modern football. This thesis offers a data-driven analysis of the importance of ball possession metrics for professional football league ranking prediction. By analysing the effects of adopting ball possession based game styles, the research seeks to show how possession-oriented strategies might help teams maximize their performance.
The algorithms used for the ranking predictions are XGBoost, LightGBM, Random Forest, Ridge Regression, and performance of all the models in predicting the goal count and possession is evaluated. The research studies the importance of ball possession metrics of both the home team and the away team in predicting the league rankings. Among the models, Random Forest model generated the lowest Mean Absolute Error values for home goal counts (0.0737), away goal counts (0.0974), and home possession (0.0289), away possession (0.0132). These results show the effectiveness of Random Forest in predicting football league rankings based on possession metrics.

1 Introduction

Football stands as one of the most globally revered sports, captivating the hearts of millions with its dynamic gameplay and fierce competition. Data analytics and innovative strategies have raised football standards in the previous decade. This thesis examines revolutionary impact of possession-based statistics for football league standings in intensely competitive English Premier League (EPL), long dominated by counter-attacking teams.

1.1 Background

Possession-based techniques have become a hallmark of great football clubs. Clubs which prioritize ball possession and retention are achieving great offensive and defensive results As a result of better game management and more scoring chances, possession-heavy teams in English Premier League often beat their opponents. Under Pep Guardiola's savvy leadership, Manchester City stands out as a prime example of this strategic strategy. Success of squad, as seen by their several league titles, is testament to efficacy of possession-based techniques. Guardiola's football philosophy is on maintaining high percentage of possession applying constant pressure. They were so dominant that other teams started concentrating on ball possession as a strategy to compete. Struggling Arsenal team hired

possession-based football coach Mikel Arteta after they failed to crack top four. In the 2022–2023 season, they finished in second place after adjusting their play style. This exemplifies how crucial it is for side to maintain control of ball in contemporary football, much like scoring goals.

1.2 Importance

This research digs into the world of data-driven football analytics, with a particular emphasis on possession-based measurements and their dramatic influence on team's performance in EPL. This study examines the vital subject of how possession-oriented tactics impact team rankings, providing useful data for teams, coaches, and analysts looking to improve their tactical methods and climb the league standings.

1.3 Research Question and Objectives

The key research questions driving this project are: What effect do ball possession-based metrics in home and away games have on forecasting football league positions? Which Machine Learning Model is best suited for predicting goal count and possession in football matches? To find the answer to these questions, the study leverages machine learning techniques such as XGBoost, LightGBM, Random Forest, and Ridge Regression.

1.4 Limitations

While this study sheds light on significance of possession-based metrics, it acknowledges certain limitations inherent to the scope of analysis. Factors such as evolving team dynamics, player injuries, and unforeseen game events may introduce complexities that the study may not fully capture. There are also other factors like change in the team ownership, management, as well as foreign investments that might affect the performance of a team.

1.5 Outline of the Structure

The next parts of the thesis unfold as follows: Chapter 2 gives a literature review to contextualize the study within existing research, while Chapter 3 shows the methodology employed, encompassing data collection, preprocessing, and the application of machine learning algorithms. Chapter 4 presents the analysis of the result and Chapter 5 is about the test cases. Chapter 6 discusses the Results and Test cases. Chapter 7 gives concluding remarks and Chapter 8 gives recommendations for the future research, and a reflection on the broader implications of possession-based strategies in contemporary football. Lastly, Chapter 9 acknowledges all the people who have helped in the completion of the thesis.

2 Literature Review

2.1 Prediction of Football Match Outcomes with Machine Learning

A variety of Machine learning algorithms have been used in predicting the results of football matches in the most efficient manner. In Cavus and Biecek (2022), a machine learning

model is developed focusing on the expected goal (xG) metric, employing tree-based classification models from XGBoost, Random Forest, LightGBM, CatBoost and explainable artificial intelligence (XAI) tools. This approach differs from the thesis by concentrating on individual shot events and their expected goal, as opposed to the broader match outcomes (goals and possession metrics) that the current thesis addresses. In Rodrigues and Pinto (2022), machine learning methods like Random Forest, XGBoost were used for forecasting football match outcomes. The study's focus on profitability of predictions when predicting the winner for betting. The current thesis and Rodrigues and Pinto (2022) both highlight the usefulness of machine learning in football match prediction but apply these techniques for different end goals.
In Baboota and Kaur (2019), a predictive model for the English Premier League is created, emphasizing feature engineering and advanced machine learning techniques. The model's performance heavily relies on main features, with the best model using gradient boosting achieving an output of 0.2156 on the probability score. The study's specific focus on league outcomes based on goals scored differs from the broader possession-based metrics employed in the thesis. In Majumdar et al. (2022), an analysis of the link between player load, injury, and match outcomes is conducted using algorithms like decision trees, logistic regression, random forest, gradient boosting. This study focuses on addressing imbalanced datasets common in injury prediction. Both the thesis and this study utilize similar machine learning algorithms, but the study emphasis on player load and injury prediction while the current model focuses on match outcomes. Study conducted by Ren and Susnjak (2022) presented unique methodology for predicting football match outcomes. This strategy included utilizing Kelly Index to categorize matches according to their level of predictability, while employing range of machine learning methods. Paper's novel utilize of Kelly Index to categorize match complexity with its focus upon comprehensible machine learning are in line with thesis's objective of using interpretable models. Present thesis model has capability to forecast both individual matches league table standings by using possession as determining factor. However, main emphasis of article was upon classifying matches according to their level of difficulty in terms of prediction.
Research conducted by Jawade et al. (2021)centers upon utilization of machine learning algorithms to forecast match results into Spanish La Liga. It incorporates various factors such as final scores, starting 11 players, substitutes, and probable goal scorers. Authors attained a precision of 71.63% through applying Logistic Regression upon database including 5 seasons. Zaveri et al. (2018) use machine learning algorithms to forecast match results in certain football leagues, having focus on identifying important characteristics and attaining high levels of precision. KNN model was determined to have superior predictions compared to SVM and Random Forest. This study enhances investigation conducted in thesis concerning use of machine learning methods for predicting football match outcomes. Present study primarily examines possession-oriented table prediction, whereas also considering particular features which impact match results.

2.2 Possession-Based Football Analytics

Football match analysis includes ball possession in various parts of the field. Bauer and Anzer (2021) use over 11,000 defensive turnover instances to automatically discover football counter-pressing techniques. Extreme gradient boosting model reveals how teams get ball back via counter pressing and how many shots they allow or produce. Counter

pressing is essential to possession, supporting the thesis's focus on possession measures. Verma et al. (2020) use 2017-2018 ball possession-position data to study English Premier League clubs' playing styles using an unsupervised k-means cluster. Unsupervised learning is used in the research to gain insights from possession statistics in different pitch zones, complementing thesis's examination of match results and league standings. The present thesis uses supervised learning, however this study uses unsupervised k-means cluster to get insights.
Chawla et al. (2017) describe a spatiotemporal trajectory data-based pass quality evaluation system. This research used computational geometry classifiers to obtain 90.2% pass rating accuracy. Pass quality is distinct from the thesis's league rankings prediction, although both use machine learning and emphasize spatiotemporal data. Gu et al. (2023) propose deep generative machine learning Possession Evaluation Model to measure team space-control efficacy. This model predicts using Premier League monitoring using event data and Variational Recurrent Neural Network (VRNN), unlike the thesis, which focused on spacecontrol efficiency. In Kusmakar et al. (2020), machine learning model segments soccer match data into events ending with goal attempts to analyze team dynamics and player interactions. The research's emphasis upon player interactions and team scoring is distinct comparing with research's emphasis on possession measures for league ranking projection..

2.3 Machine Learning for Ranking Prediction

Football analytics involves predicting league or tournament winners. Tax and Joustra (2015) use public data as well as reduction of dimensionality to create a Dutch Eredivisie match forecasting system which can be used to predict the ranking of Everivisie teams. Naive Bayes or Multilayer Perceptron integrated with PCA was found to be the best models. The study also explores a hybrid model incorporating odds of the winner, relevant to the thesis in its use of different machine learning techniques and the impact of dimensionality reduction on prediction accuracy. In Pantzalis and Tjortjis (2020), advanced statistics and historical data are leveraged to predict final league tables for selected leagues. This aligns with the thesis's objectives of using machine learning algorithms for league ranking and possession metrics prediction, but different models were used in the current model and the training was done in Premier League dataset.
In Radhika and Syed Masood (2022), feature selection in football analytics is of prime importance. Logistic Regression, Random Forest, Multinomial Naïve Bayes, SVM are trained and tested. Logistic Regression achieves over 80% accuracy in the English Premier League, aligning with the thesis in its use of Random Forest, logistic regression models. However, the thesis extends the application to specifically correlate ball possession metrics with league standings. Both the study and the thesis emphasize the main role of feature selection in increasing the accuracy of predictions in football analytics, though the thesis has a focused application in associating ball possession with league performance.
In Groll et al. (2021), statistical ranking methods are combined with a hybrid machine learning model to forecast UEFA EURO 2020, employing a Random Forest model. The model predicted the Euro 2020 winners as France National Football Team with a probability of 14.8%. While the thesis centers on predicting league rankings, this study aims at forecasting tournament outcomes using various data, including socio-economic factors.
In Li et al. (2020), a Linear Support Vector Classifier model is applied to rank teams in

the Chinese Football Super League, achieving a high correlation with actual league standings. Both the study and the thesis demonstrate the effectiveness of data-driven models in football outcome prediction, but with different focal points—the study on overall team performance and the thesis on the influence of ball possession on league standings.

2.4 Football Analytics Based on Bookmaker Predictions

Several new betting platforms have arisen in previous decade, highlighting crucial significance of properly forecasting match outcomes in this industry. Research conducted in Esme and Kiran (2018) investigates use of betting odds in conjunction with KNN algorithm for purpose of predicting football match outcomes. This algorithm assesses the similarities between contests by analyzing betting patterns and incorporates statistics on performance from prior games to improve accuracy. This is consistent with main aim of the thesis, which is to use data-driven methodologies in football analytics. It provides additional strategy to anticipate match results. Study conducted by Odachowski and Grekow (2013) examines effectiveness of bookmaker odds in predicting outcomes of football matches. Classifiers use bookmaker odds variations for forecasting match results by developing link amongst these changes. Research demonstrated a 70% efficacy in making predictions, indicating substantial association between fluctuations in bookmaker odds and results of matches. These findings imply the possibility of constructing a decision-making system which can forecast outcomes by analyzing fluctuations in chances. This research enhances thesis by offering an alternate viewpoint utilizing bookmaker odds.
Egidi et al. (2018) developed Bayesian Poisson model that integrates past data and betting odds to forecast the results of football matches. The study focuses on the significance of predicted accuracy in statistical models, particularly within a Bayesian framework. Process entails using model and posterior estimates to simulate matches while offering predictions, such as final rankings and probability of positions. Study conducted by Stübinger et al. (2019) investigates use of machine learning in forecasting football league results. It emphasizes need of a thorough feature set and human supervision in predicting models. This human error may be deduced from past betting trends. Research places significant importance upon many aspects that influence results of matches, providing comprehensive perspective that includes consideration of human decision-making, in contrast to thesis's narrow concentration just on ball possession numbers. Esumeh (2015) examines utilization of machine learning in making predictions for football betting. Tree-based algorithms like Random Forest forecast football matches better. Complex predictive models provide better forecasts and greater average payoffs. The study shows that an ensemble strategy of machine learning algorithms can produce statistically and economically significant returns, similar to the thesis's use of advanced machine learning for football predictions but focusing on betting strategies.

3 Methodology

This section outlines the structured methodology employed in conducting the research on the influence of possession-based metrics in predicting football league rankings, specifically focusing on home and away games within the English Premier League (EPL) in the past five seasons. The methodology is designed to address important aspects of the study. This section includes data collection, preprocessing, and applying advanced ML algorithms.

3.1 Data Selection and Data Segregation Process

The datasets chosen for this study is the officially available data of the Premier League matches played in the past five seasons from 2018/2019 season to 2022/2023 season. The dataset was obtained from a publicly available link[1] available on the internet.

For predicting the winners of the 2023/2024 season, the Premier League data from 2017 to 2023 is used. Python includes the Pandas library. DataFrames with crucial football match and team information are merged during data merging using Pandas. This operation creates two CSV files, "match_data.csv" and "teams data.csv". All match and team data by 5 seasons evaluated are integrated in 2 files, "match data.csv" and "teams_data.csv".
Football match data is consolidated in DataFrame in "match data.csv" file. This DataFrame has 1900 rows, each indicating an exceptional match, and 66 columns, every representing a feature. Goal statistics, possession percentages, card distributions, and other information are included. Produced CSV file provides a thorough and structured dataset for match analysis of performance over seasons.
Meanwhile, code mixes team-specific DataFrames for team statistics. The synthesized information is then stored in the teams_data.csv file. This file exhibits a structured format with dimensions (100, 293), signifying 100 rows, each representing a distinct team, and 293 columns, encapsulating various team-related features. The amalgamated team statistics provide a rich dataset that can be harnessed for analytical purposes, offering insights into team performances based on an extensive array of metrics. These CSV files, generated through the Data Merging process, stand as valuable repositories of football-related data.
1

3.2 Data Preprocessing and Data Visualization

Data Preprocessing demonstrates data exploration and manipulation using the pandas library in Python in such a way that only the data that is required is used for model training. The missing values are found out and deleted. Next, the data types like object are converted to useful data types for further analysis. From among 66 columns, only the columns named ' hometeamname', 'hometeamgoalcount', 'hometeamcornercount', 'hometeamshots', 'hometeamshotsontarget', 'hometeamshotsofftarget', 'hometeampossession', 'teamaxg', 'totalgoalcount','awayteamname', 'awayteamgoalcount', 'awayteamcornercount', 'awayteamshots', 'awayteamshotsontarget', 'awayteamshotsofftarget', 'awayteampossession', 'teambxg'are selected and saved to preprocessed_data.csv. This data contains only the required details about home team and away team goals and possession, whichis required for building the model.

The above Figure 1 represents bar graph, illustrates the number of samples collected for a football match over the years. In 2018, there were 200 samples, followed by an increase to 379 in 2019. The number decreased to 336 in 2020 and rose to 408 in 2021, decreased again to 361 in 2022. The data for 2023 shows a further decrease to 216 samples. Overall, the graph reflects variations in sample collection across the specified years.

A countplot to visualize the distribution of total goal counts is given in the above Figure 2. The above plot has peaks at 2 and 3, which means that the most common

[1] `https://footystats.org/england/premier-league`

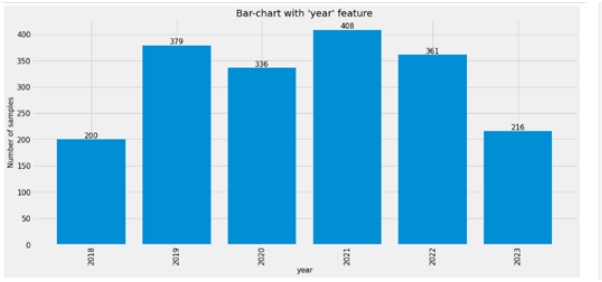

Figure 1: Bar Chart showing matches played in a year

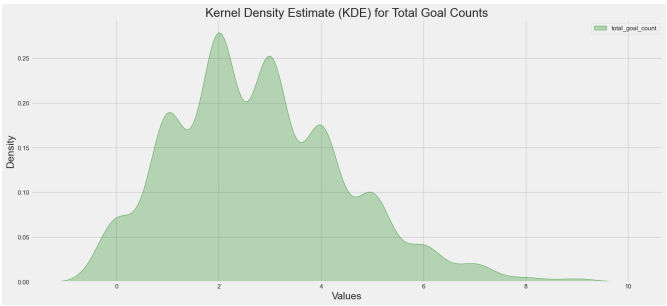

Figure 2: Kernel Density Estimate of Total goal count

goal line scores are either 1-1, 2-0, 3-0, 2-1 , 1-2, 0-3 etc. However, due to home team advantage, the most likely scores are 1-1, 2-0, 2-1 and 3-0.
From the "teams_data.csv" file, we get the number of teams that have played all the five seasons and the teams that got relegated during this period.

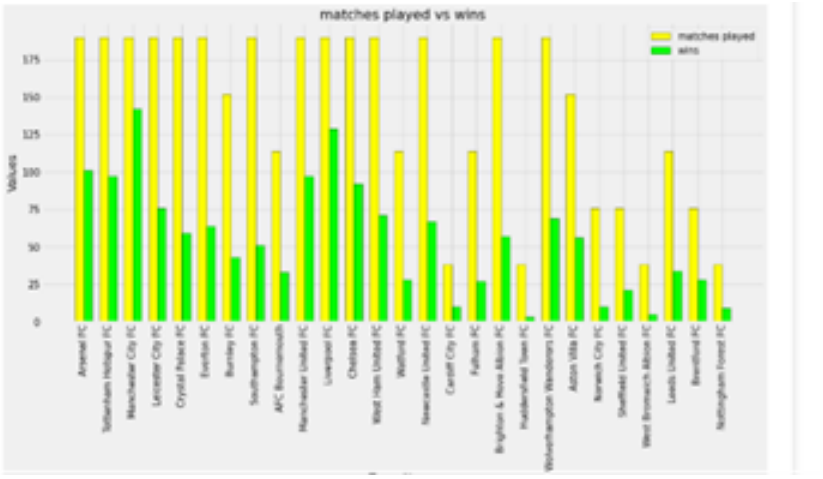

Figure 3: Matches played vs Wins of Premier League Teams

From the above Figure 3, from the matches played, we can see that only the teams that played the whole 5 seasons are Arsenal, Manchester City,Leicester City FC, Crystal Palace, Everton , Tottenham Hotspur , Southampton, Manchester United FC Liverpool FC, Chelsea, West Ham United, Newcastle United, Brighton and Hove Albion FC, Wolverhampton Wanderers FC. All the other teams have got relegated from the top flight

at some point between 2017 to 2023. The teams with the best win percentage over the 5 seasons are Manchester City FC, Liverpool FC, Arsenal FC, Manchester United FC, Tottenham Hotspur FC, Chelsea FC.

As a result of the pre-processing step, we now have the "preprocessed_data.csv" file with the dimensions (1900, 17) which will be used for the further in the analysis.

3.3 Model Selection

Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBRegressor, LGBMRegressor, RandomForestRegressor and Ridge are among the Python programming libraries utilized in this work for data analysis and modelling. The training dataset consists of football match data from official Premier League databases, including possession metrics, game results, and numerous performance related information.

3.3.1 XGBoost

XGBoost, also known as Extreme Gradient Boosting, is gradient boosting method that is highly regarded for its exceptional speed and efficiency. It efficiently manages intricate and extensive datasets, making it well-suited for predictive modeling. XGBoost algorithm monitors football game data in sports analytics, including possession proportions, shot counts for home as well as away games, etc. Ball possession metrics are important in football outcomes because they show complex correlations between this data and league rankings.

3.3.2 LightGBM

LightGBM, or Light Gradient Boosting Machine, is quick efficient gradient boosting system for huge datasets. XGBoost develops more complex trees with leaf-wise split technique, that might enhance accuracy but increase overfitting. LightGBM is known in football statistics for its fast and accurate possession data analysis, which helps explain team rankings. Possession-focused football league standings forecasting algorithms need tool's capacity to effectively analyze big datasets using minimal training time.

3.3.3 Random Forest

Combining decision tree outcomes enhances accuracy and decreases overfitting in Random Forest ensemble learning. Random Forest handles complex datasets well. Football analytics employs RF's ensembles technique for analyzing complex football data to illustrate how possession-based strategies affect team standings across games. Versatility and extensive study assist football predicting methods.

3.3.4 Ridge Regression

Ridge Regression uses regularization to prevent overfitting convergence. Ridge Regression works effectively with tightly connected predictor variables. Ridge Regression is employed in football statistics as it combines model flexibility with prediction accuracy, making it vital for analyzing complex possession data-club ranking correlations. Analytical frameworks promote interpretability and generalization. This makes it beneficial for understanding football league standings dynamics.

3.4 Model Training

The selected machine learning models, XGBoost, LightGBM, Random Forest, Ridge Regression, are trained to predict football match outcomes. The home team goal count, away team goal count, home team possession, and away team possession, are predicted during this phase of the research. With train_test_split function with a test size of 20% and random state of 42, dataset acquired from the preprocessed_data.csv file was suitably dividedinto training and testing sets.

3.4.1 Training Goal Count Prediction Model of Home Team and Away Team

This approach uses two independent algorithms to forecast goal totals for the home and away teams. These modules employ machine learning techniques to forecast the number of goals a home team / away team will score in a football match. This dataset, extracted from a CSV file, encompasses diverse features such as team names, goal counts, corner counts, shots, possession percentages, and expected goals (xG). Data is split in training and test sets for model assessment. The research predicts home and away team goal counts using four algorithms: Random Forest Regressor, XG Boost, LGBM and Ridge. This phase trains eight models—four for home goal count and four for away goal count.

3.4.2 Training Possession Prediction of Home Team Possession and Away Team

This technique utilizes two distinct models to forecast team possession for both home team and the away side. The purpose of this program is to function on preprocessed_CSV file that contains relevant characteristics related to football matches, such as team names, goals counts, possession % etc. The data undergoes a division to training and testing set to facilitate the application of regression algorithms, specifically Random Forest Regressor, XG Boost Regressor, LGBM Regressor, and Ridge for possession prediction of both home and away teams. Thus, a total of 8 models are trained in this phase, four models for predicting home team possession and four models for predicting away team possession.

3.5 Model Evaluation

The Random Forest Regressor undergoes training on the designated training set and subsequently undergoes evaluation on the test set. The predictions are then converted to integer values. To facilitate future use, the trained model is stored using the pickle module. Performance of this analytical models which predicts goal count and possession are studied using key metrics, includingRoot Mean Squared Error (RMSE), the Mean Squared Error (MSE), Mean Absolute Error (MAE). These metrics are evaluated as follows:

- Mean Absolute Error (MAE): The average absolute difference between the actual and anticipated values is represented by MAE. It gives a simple way to assess the model's predicting accuracy. Lower MAE suggests improved performance.
- Mean Squared Error (MSE): MSE is a calculation that takes the average of the squared differences between the actual and forecasted values. Squaring the differences penalizes greater mistakes more severely. A lower MSE, like a lower MAE, is desired.

- Root Mean Squared Error (RMSE): The square root of the MSE is the RMSE. It has the same scale as the target variable, making it easier to comprehend. It also penalizes larger mistakes more severely and offers information about the predicting accuracy of the model.

Likewise, the XGBoost Regressor, LGBM Regressor and Ridge Regressor models are implemented, and their predictive capabilities are scrutinized using the same evaluation metrics. The trained models are saved as pickle files so that they can be invoked to predict the results for various test cases.

3.6 System Architecture

For football match predictions based on possession-based metrics, the system design is a two-phase model training and prediction system. The approach uses a historical football dataset to train four machine learning models: two models to predict goal counts for both teams and two models to predict possession for both home and away sides. A diagram of system architecture is shown in Figure 4.

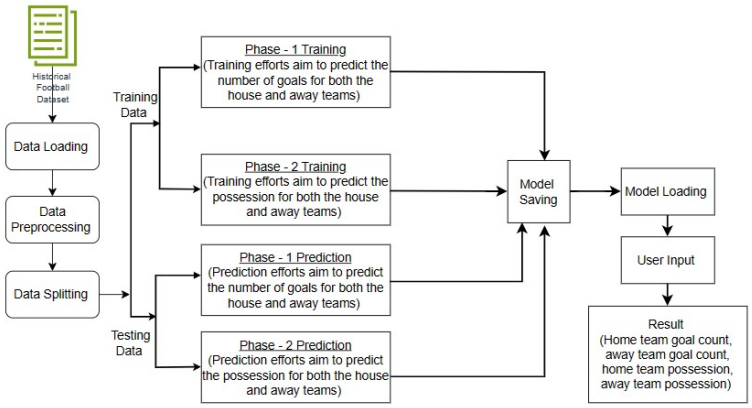

Figure 4: System Architecture of the Model

Phase 1: Training The system loads and preprocesses the historical football dataset in the first step. This might entail cleansing the data and transforming it to a consistent format. After preprocessing, the data is divided into training and testing sets. The machine learning models are trained using the training set. System employs a supervised learning technique, which means it feeds the models input data as well as the intended output (e.g., number of goals scored, possession percentage). Based on the input data, the models learn to predict the output.
Phase 2: Prediction Once the models are trained, they are used to predict the result of new football matches. Goal count and possession statistics for match are fed in models. Now, models predict both teams' goals and possession percentages.
Advantages of this algorithm are

- Adaptability: System is adaptable to football leagues and tournaments. This is because it is independent of qualities and data.

- Precision: The technology can predict match results quite well. It uses two machine learning models to predict game factors like goals and possession.

- Scalability: The system is able to handle large datasets and significant number of predictions in an effective manner. The reason for it is because it makes utilize machine learning models that have been especially designed for scalability.

With aim of forecasting football matches, this system architecture offers solution which is robust and flexible.

4 Results

This section examines the effectiveness of algorithms for forecasting goal counts and possession. We assessed total of 16 models, in 8 models specifically designed for predicting goal counts and eight additional models focused on predicting ball possession.

4.1 Performance evaluation of Models predicting goal count of Home Team and Away Team

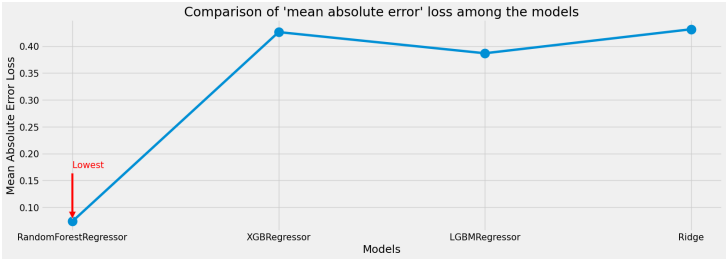

Figure 5: MAE comparison for Home Team goal count

Random Forest Regressor model demonstrates a MSE of 0.074, RMSE of 0.271, and MAE of 0.0737 in predicting home team goals. XGBoost Regressor model produces errors with MSE of 0.426, RMSE of 0.653, and MAE of 0.4263. The LGBM Regressor yields an MSE of 0.387, RMSE of 0.622, and MAE of 0.3868. Finally, the Ridge model produces predictions having an MSE of 0.432, RMSE of 0.657, and MAE of 0.4316. Present investigation demonstrates practical use of machine learning algorithms in predicting the number of goals scored by home team in football matches. Based on error metrics, evaluation concludes that Random Forest Regressor outperforms other three algorithms with an MAE of 0.0737. This highlights its effectiveness in making predictions in this context. Figure 5 depicts the MAE comparison of all the above models.

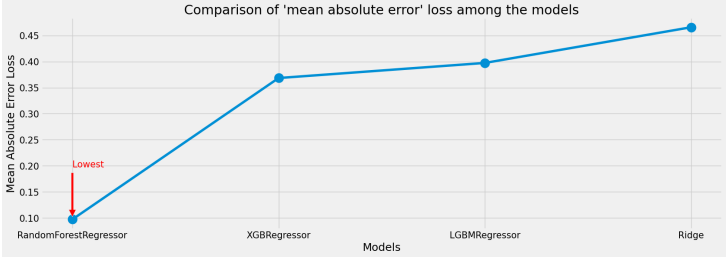

Figure 6: MAE comparison for Away Team goal count

For away team goal prediction also the Random Forest Regressor model exhibits a mean squared error (MSE) of 0.097, root mean squared error (RMSE) of 0.312, and mean absolute error (MAE) of 0.0974. The XGBoost Regressor model yields errors with MSE of 0.368, RMSE of 0.607, and MAE of 0.368. Similarly, the LGBM Regressor produces MSE of 0.397, RMSE of 0.630, and MAE of 0.3974. Lastly, the Ridge model results in predictions with MSE of 0.466, RMSE of 0.682 and MAE of 0.4658. Random forest regressor, with the least MAE of 0.0974 is the most efficient model among the four models studied. Figure 6 depicts the MAE comparison of all the above models.

4.1.1 Performance evaluation of Models predicting ball possession of Home Team and Away Team

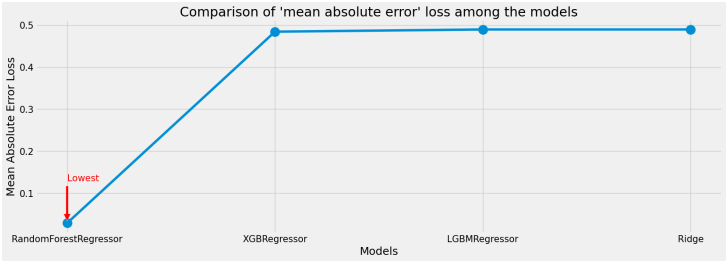

Figure 7: MAE comparison for Home Team goal possession

For prediction of team possession of the home team, the Random Forest Regressor model exhibits a mean squared error (MSE) of 0.029, root mean squared error (RMSE) of 0.170, and mean absolute error (MAE) of 0.0289. The XGBoost Regressor model yields higher errors with MSE of 0.484, RMSE of 0.696, and MAE of 0.4842. Similarly, the LGBM Regressor produces MSE of 0.516, RMSE of 0.718, and MAE of 0.4895. Lastly, the Ridge model results in predictions with MSE of 0.489, RMSE of 0.700, and MAE of 0.4895. These metrics assess the models' performance by comparing their predictions to the true values. Random Forest Regressor, with the lowest MAE score of 0.0289 is found to be the best model in predicting ball possession of home teams. Figure 7 depicts the MAE comparison of all the above models.

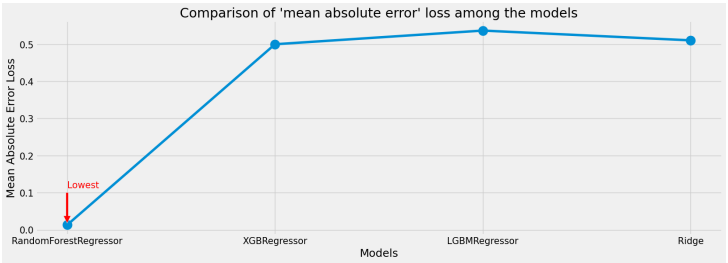

Figure 8: MAE comparison for Home Team goal possession

Similarly, the prediction for the away team the Random Forest Regressor, the Mean Squared Error (MSE) is 0.013, Root Mean Squared Error (RMSE) is 0.115, and Mean Absolute Error (MAE) is 0.0132. The trained model is saved using pickle. Similarly,

XGBoost Regressor achieves an MSE of 0.500, RMSE of 0.707, and MAE of 0.5000. The model is saved as well. LGBM Regressor yields an MSE of 0.558, RMSE of 0.747, and MAE of 0.5368. Finally, the Ridge algorithm produces predictions with MSE of 0.511, RMSE of 0.715, and MAE of 0.5105. Random Forest Regressor with MAE of 0.0132 is found to be the best suited model for predicting away team possession. Figure 8 depicts the MAE comparison of all the above models.

5 Test Cases

The models trained and stored using pickle module can be used to generate test cases which can further be used to predict the winner of individual matches based on user input and also predict the league table ranking in any particular Premier League season. The inferences received from these test cases can be used to get more information regarding the efficacy of the possession metrics related prediction model.

5.1 Predicting the Top 6 table of 2021/2022 Premier League Season with emphasis on possession

In the field of football analytics, a thorough investigation of the predictive potential of ball possession measurements for establishing Premier League rankings unfolded over the course of a careful research spanning the seasons 2016/2017 to 2021/2022. The data from 2019/2020 season is omitted as the Premier League matches were effected due to Covid in that season. The study attempted to predict home and away team goal results as well as possession percentages using complex machine learning models, especially the Random Forest Regressor and the XGBoost Regressor. These models are loaded with the help of pickle function. The primary goal was to show the significant effect of possession-centric football methods in a team's points table rankings.
The study attempted to build similarities between projected and real top 6 tables by assigning point values to clubs based on expected goal differentials and possession percentages under different match situations. This analysis highlighted the critical importance of possession in determining football outcomes. When estimating the table, the following conditions were taken into account:

1. If the expected home team goal exceeds the predicted away team goal and the predicted home team possession exceeds the predicted away team possession, the home team earns three points and the away team receives none.

2. If the expected home team goal is less than the predicted away team goal, and the predicted home team possession is less than the predicted away team possession, the home team receives 0 points while the away team receives 3.

3. When the expected home team goal exceeds the predicted away team goal but the predicted home team possession falls short of the predicted away team possession, both teams receive one point.

4. If, on the other hand, the expected home team goal is lower than the predicted away team goal, but the forecast home team possession surpasses the predicted away team possession, both teams earn one point.

	Rank	Actual Top 6 (2022)	Top 6 (Random Forest)	Variation (Random Forest)	Top 6 (XGBoost)	Variation (XGBoost)
	1	Manchester City	Manchester City	0	Manchester City	0
	2	Liverpool	Liverpool	0	Liverpool	0
	3	Chelsea	Chelsea	0	Tottenham Hotspur	+/-1
	4	Tottenham Hotspur	Tottenham Hotspur	0	Chelsea	+/-1
	5	Arsenal	Manchester United	+/-1	Manchester United	+/-1
	6	Manchester United	Arsenal	+/-1	Arsenal	+/-1

Table 1: Comparison of Actual and Predicted Top 6 Teams 2021/2022 season

We get the above Table 1 after running both the models. This table vividly showcases the alignment and disparities between actual and predicted top 6 tables. Both the table shows the same top 6 teams, but the rankings vary in both. Both the models show the first and the second positions accurately. The Random Forest Model predicts the position of four of the six teams accurately, while XGBoost predicts two of the six teams accurately, with variations coming in small +/-1 positions. This table reinforces the thesis's assertion on the decisive role of possession in shaping football league rankings.

5.2 Case Study 2: Predicting the Top 6 table of 2023/2024 Premier League Season with emphasis on possession

In Table 2, the winners of the 2023/2024 season is predicted. For this, we train the models with data from 2017/2018 season to 2022/2023 season. Similar to the above case study, the conditions are such that points are assigned to teams based on predicted goal differentials and possession percentages in specific match conditions. The below table was obtained after running both the Random Forest and XGBoost Model. With test cases, models were evaluated thoroughly. 4 of top six teams were predicted correctly by Random Forest Model and 2 by XGBoost. Both Models properly predicted the top 6 clubs with a +/-1 league ranking shift. The concept gave possession greater weight in tallying points for clubs, yet the anticipated Premier League rankings matched the actual results. This indicates that football league rankings are influenced by possession. Most of the teams that are aiming to climb from the bottom half of the table to European positions (positions from 1-7), can thus concentrate on possession based tactics to get better results.

Both the models have predicted the same top 3 teams – Manchester City, Liverpool and Chelsea. As noted from Case 1, there might be a +/- 1 difference in position from the actual data. Outliers that may arise due to change in management of the club or point deduction due to Premier League Financial Fair play regulations are not incorporated in this model. As a result, the recent management changes might affect the position of Chelsea in the below table.

Rank	Random Forest	XGBoost
1	Manchester City	Manchester City
2	Liverpool	Liverpool
3	Chelsea	Chelsea
4	Arsenal	Tottenham Hotspur
5	Manchester United	Arsenal
6	Tottenham Hotspur	Manchester United

Table 2: Predicted Top 6 Teams 2023/2024 season

5.3 Case Study 3: Predicting the winner of each match in a Premier League season

As a part of the thesis, we have developed a model to forecast the results of each match based on possession metrics. This model shows the list of available home and away teams, and user can input the teams of his choice. Once the user inputs the teams, the model will predict the winner with the help of the models already trained and stored in the pickle file.
Since, the highest accuracy was received for the Random Forest model from our studies conducted above, we have used the goal count and possession predicted by the Random Forest model in the match winner prediction. Once the Goal count and possession stats are predicted, the average of predicted home and away goal count and possession is calculated. These four values obtained from four different Random Forest models are compared and computed to predict the winner based on ball possession metrics. The same conditions in Case 1 were used to predict the winner. When predicted home team goal count and away team goal is greater than away team goal and possession, the home team is considered the winner. When predicted away team goal count and home team goal is greater than home team goal and possession, the away team is considered the winner. In all the other cases, the match result is concluded to be a draw.
Three different test cases are studied below. Goal(P), Possession(P), Winner(P) shown in the tables are the predicted goal count, possession and winner of the match:

5.3.1 Predicting Manchester United vs Chelsea score in 2021/2022 season

Home/Away Team	Goal(P)	Goal	Possession(P)	Possession	Winner	Winner(P)
Manchester United	1	1	48	35	Draw	Draw
Chelsea	0	1	52	65	-	-

Table 3: Predicted vs actual winner of Manchester United vs Chelsea (2021/2022)

Table 3 depicts the expected and actual outcome of a match between Manchester United and Chelsea. The predicted goal count is 1 for Manchester United and 0 for Chelsea. This shows that if only the predicted goal count was considered, the model would have predicted that Manchester United will win the game. However, since the predicted ball possession of Chelsea was higher than Manchester United, our model predicted that

Manchester United might not dominate the game and the match might end in a draw. Hence, with the ball possession metrics-oriented prediction, the result was predicted accurately as a draw.

5.3.2 Predicting Manchester City vs Everton score in 2021/2022 season

Home/Away Team	Goal(P)	Goal	Possession(P)	Possession	Winner	Winner(P)
Manchester City	2	1	68	77	Manchester City	Manchester City
Everton	1	1	32	23	-	-

Table 4: Predicted vs actual winner of Manchester City vs Everton(2021/2022)

Table 4 depicts the expected and actual outcome of a match between Manchester City and Arsenal. Here we can see that the model has predicted more goals for Manchester City and also indicates that they will dominate possession of the ball by 68 percent. This shows that Manchester City has more possibility to dominate the game against Everton and match wont be very competitive. The actual result also aligns with the above predicted values and the winner of the match is predicted accurately.

5.3.3 Predicting Tottenham vs Fulham winner in 2023/2024 season (Already concluded match in current season)

Home/Away Team	Goal(P)	Goal	Possession(P)	Possession	Winner	Predicted Winner
Tottenham	2	2	54	56	Tottenham	Tottenham
Fulham	1	0	46	44	-	-

Table 5: Predicted vs actual winner of Tottenham vs Fulham (2023/2024)

Table 5 shows the predicted and actual result in a match between Tottenham and Fulham in the 2023/2024 season. The predicted goal count and possession aligns with the actual result, proving the applicability off the model in the current Premier League season. The actual winner in the game was correctly identified as Tottenham using the model.

5.3.4 Predicting the Liverpool vs Manchester City winner 2023/2024 season (Future game)

Home/Away Team	Goal(P)	Possession (P)	Predicted Winner
Liverpool	1	47	Draw
Manchester City	1	53	-

Table 6: Predicted winner of Liverpool vs Manchester City(2023/2024)

Table 6 shows the predicted result in a match between Liverpool and Manchester City that has not been concluded yet. These teams were selected in this case study as they are the top 2 teams in the league based on win percentages and hence, the result of this game garners a lot of interest among football fans. The possession-metrics based model predicts that this match will conclude in a draw. It also infers that the match

will be tightly contested as both the teams have comparable possession. This is a metric very important for bookmakers as the model predicts a high level of competitiveness, the stakes will be higher for betting and right predictions will lead to higher gains for the bookmakers.

6 Discussion

The thesis predicts Home Team goal count, Away Team goal count, Home Team possession and Away Team possession with each of the above predicted by four different models: Random Forest, XGBoost, LightGBM and Ridge regression model. Thus, a total of 16 Models were trained in this project work. Random Forest with an MAE score of 0.0737, 0.0974, 0.0289, 0.0132 predicted for Home Team goal count, Away Team goal count, Home Team possession and Away Team possession respectively was found to be the best suited model for ball possession metrics oriented prediction. XGBoost with an MAE score of 0.4263, 0.368, 0.4842, 0.5000 for the corresponding metrics above is adjudged to be the next best model for prediction.
The Random Forest model was able to outperform the other models due a number of reasons. The Model's ability to handle complicated, non-linear connections found in sports data is a huge benefit. It can better forecast football league rankings with this capacity. Random Forest's ensemble method prevents overfitting, a significant issue with Ridge regression. Its accurate predictions of Home and Away goal counts and possessions are likely due to its robustness. Football statistics often incorporate category and numerical data, making Random Forest beneficial. It may analyze sports data without preprocessing, making it good option. Finally, Random Forest excels with huge datasets like historical football data. Model's capacity to manage and operate with large datasets is important to its success.
With test cases, models were evaluated thoroughly. 4 of top six teams were predicted correctly by Random Forest Model and 2 of the top 6 teams were predicted correctly by XGBoost. Both Models properly predicted the top 6 clubs with a +/-1 league ranking shift. The concept gave possession greater weight in tallying points for clubs, yet the anticipated Premier League rankings matched the actual results. This indicates that football league rankings are influenced by possession.
In the test situations, the model properly predicted outcomes of completed matches. This model also predicted outcome of match among top two teams. Model anticipated both outcome and match's competitiveness, key bookmaker statistic.

7 Conclusion

A total of four machine learning models: Random Forest, XGBoost, LightGBM and Ridge regression model were trained and tested with the Premier League dataset to predict goal count and possession. The algorithm best suited for predicting home and away goal count and possession was found to be Random Forest with an MAE score of 0.0737, 0.0974, 0.0289, 0.0132 for Home Team goal count, Away Team goal count, Home Team possession and Away Team possession respectively. With the help of these predicted values, the ball possession metrics-oriented model predicted the league rankings of the top 6 teams in the Premier League with a maximum league standing position change of +/-1. This shows that possession is an important metric when predicting league standings in football.

Coaches, teams, and analysts can use these insights to refine their tactics, hereby focusing more on ball possession as a critical factor for success.
Team that has more predicted goal count and predicted possession indicates that the team has dominated the match against their opponents. This shows a higher certainty in predicting the winner than solely using the predicted goal count to predict the winner as in many cases, the goal scored might have been against the run of play. Thus, the prediction proposed by the ball possession metric-oriented model has more credibility in predicting a football match than predicting a match using a model based on just the goal count.

8 Future Work

The thesis significantly contributes to advancing predictive modelling in football analysis, bridging the gap between possession strategies and goal outcomes. Future work may involve refining models with additional features like possession in spaces with higher xG for scoring goals, exploring real-time predictions which can incorporate factors like red cards, injuries to key players, and adapting the system for different leagues. Furthermore, integrating other variables such as player form and fitness, weather circumstances, and psychological considerations into football match forecasts might provide a more holistic approach in the future.

9 Acknowledgement

I would like to express my sincere gratitude to Sasirekha Palaniswamy for her supervision, meticulous guidance and support throughout the research. Her patience, motivation and immense knowledge helped me at the time of the research. I would like to thank my parents, sister, colleagues and friends for their support and motivation during the thesis work.

References

Baboota, R. and 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.

Bauer, P. and Anzer, G. (2021). Data-driven detection of counterpressing in professional football: A supervised machine learning task based on synchronized positional and event data with expert-based feature extraction, *Data Mining and Knowledge Discovery* **35**(5): 2009–2049.

Cavus, M. and Biecek, P. (2022). Explainable expected goal models for performance analysis in football analytics, *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 1–9.

Chawla, S., Estephan, J., Gudmundsson, J. and Horton, M. (2017). Classification of passes in football matches using spatiotemporal data, *ACM Transactions on Spatial Algorithms and Systems (TSAS)* **3**(2): 1–30.

Egidi, L., Pauli, F. and Torelli, N. (2018). Combining historical data and bookmakers' odds in modelling football scores, *Statistical Modelling* **18**(5-6): 436–459.

Esme, E. and Kiran, M. S. (2018). Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm, *International Journal of Machine Learning and Computing* **8**(1): 26–32.

Esumeh, E. O. (2015). Using machine learning to predict winners of football league for bookies, *Int. J. Artif. Intell* **5**: 22.

Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schauberger, G., Van Eetvelde, H. and Zeileis, A. (2021). Hybrid machine learning forecasts for the uefa euro 2020, *arXiv preprint arXiv:2106.05799* .

Gu, C., De Silva, V. and Caine, M. (2023). A machine learning framework for quantifying in-game space-control efficiency in football, *Knowledge-Based Systems* p. 111123.

Jawade, I., Jadhav, R., Vaz, M. J. and Yamgekar, V. (2021). Predicting football match results using machine learning.

Kusmakar, S., Shelyag, S., Zhu, Y., Dwyer, D., Gastin, P. and Angelova, M. (2020). Machine learning enabled team performance analysis in the dynamical environment of soccer, *IEEE access* **8**: 90266–90279.

Li, Y., Ma, R., Gonçalves, B., Gong, B., Cui, Y. and Shen, Y. (2020). Data-driven team ranking and match performance analysis in chinese football super league, *Chaos, Solitons & Fractals* **141**: 110330.

Majumdar, A., Bakirov, R., Hodges, D., Scott, S. and Rees, T. (2022). Machine learning for understanding and predicting injuries in football, *Sports Medicine-Open* **8**(1): 1–10.

Odachowski, K. and Grekow, J. (2013). Using bookmaker odds to predict the final result of football matches, *Knowledge Engineering, Machine Learning and Lattice Computing with Applications: 16th International Conference, KES 2012, San Sebastian, Spain, September 10-12, 2012, Revised Selected Papers 16*, Springer, pp. 196–205.

Pantzalis, V. C. and Tjortjis, C. (2020). Sports analytics for football league table and player performance prediction, *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*, IEEE, pp. 1–8.

Radhika, A. and Syed Masood, M. (2022). Premier league table prediction using machine learning algorithms, *Webology* **19**(1): 6379–6395.

Ren, Y. and Susnjak, T. (2022). Predicting football match outcomes with explainable machine learning and the kelly index, *arXiv preprint arXiv:2211.15734* .

Rodrigues, F. and Pinto, Â. (2022). Prediction of football match results with machine learning, *Procedia Computer Science* **204**: 463–470.

Stübinger, J., Mangold, B. and Knoll, J. (2019). Machine learning in football betting: Prediction of match results based on player characteristics, *Applied Sciences* **10**(1): 46.

Tax, N. and Joustra, Y. (2015). Predicting the dutch football competition using public data: A machine learning approach, *Transactions on knowledge and data engineering* **10**(10): 1–13.

Verma, P., Sudharsan, B., Chakravarthi, B. R., O'Riordan, C. and Hill, S. (2020). Unsupervised method to analyze playing styles of epl teams using ball possession-position data, *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, pp. 58–64.

Zaveri, N., Shah, U., Tiwari, S., Shinde, P. and Teli, L. K. (2018). Prediction of football match score and decision making process, *International Journal on Recent and Innovation Trends in Computing and Communication* **6**(2): 162–165.