Leeds University Business School

UNIVERSITY OF LEEDS

Dissertation/Project Coversheet

Student ID Number:	2	0	1	6	7	8	0	7	6
Student Name	Abhishek .								
Module Code:	LUBS5579M								
Programme of Study:	MSc Business Analytics and Decision Sciences								
Supervisor:	Dr. Sajid Siraj								
Title:	Quantitative Analysis of Footballer Performance in the English Premier League using Machine Learning Algorithms.								
Declared Word Count:	9798 words								

Please Note:

Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such. If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.

It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices ("appendix abuse"). It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams.

By submitting an assignment you confirm you have read and understood the University of Leeds **Declaration of Academic Integrity** (http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf).

Acknowledgement

First and foremost, I would like to express my profound gratitude to God Almighty for bestowing upon me strength, wisdom, and perseverance throughout the course of this dissertation period.

To my parents, your unwavering support and encouragement helped me through the most challenging times. Your faith in me has been a constant source of motivation, and for that, I am eternally grateful.

I must extend my heartfelt gratitude to Dr. Sajid Siraj for his invaluable guidance, patience, and expertise. Serving as my dissertation supervisor, his insights and constructive feedback have been instrumental in shaping this research and has enriched my experience immensely.

In this journey of knowledge and discovery, the contributions of each one of you have been monumental. I am deeply thankful for the blessings and guidance I have received throughout this endeavour.

Quantitative Analysis of Footballer Performance in the English Premier League using Machine Learning Algorithms.

Abstract

This dissertation analyses footballers performances based on their playing position in the English Premier League (EPL) for the 2022-2023 season. The primary metric under scrutiny for this study is the Average Match Rating. This research aims to discern the Key Performance Indicators (KPIs) that significantly influence this rating based on each position, while acknowledging that the determining factors of performance vary across different roles on the pitch.

To achieve the objectives, two prominent Machine Learning algorithms- Multiple Linear Regression and Random Forest, were utilized. These models were trained to predict the Average Match Rating of players based on the KPIs identified earlier in the study. This subsequently generates a ranked list of players. This derived ranking from the models was compared against the original Average Overall Match Ratings rankings to gauge the efficacy of the utilized predictive models.

The study further delves into a comparative analysis of the two machine learning algorithms by evaluating their Mean Absolute Errors (MAE) for each position. This concludes that Multiple Linear Regression must be used for evaluating the Average Match Rating for Forwards and Midfielders and Defenders while Random Forest must be used to evaluate the Average Match Rating for Goalkeepers.

Keywords: Average Match Rating, Key Performance Indicators, Multiple Linear Regression, Random Forest, Mean Absolute Error

Table of Contents

Title: Quantitative Analysis of Footballer Performance in the English Premier League using Machine Learning Algorithms.

1. Introduction

This section contains a brief introduction about the sport of football on the whole, the existing major football leagues globally, the reasons for selecting the English Premier League as the topic of study, the purpose of this research, and the research questions.

1.1 Background

Football, which is also known as soccer in various parts of the world, is a team sport in which two teams of eleven players each compete. It can be traced back over 2,000 years to ancient civilizations including China, Greece, and Rome. (Goldblatt, 2008). The modern version of the game, however, began in England in the 19th century. The most popular sport in the world is football (Palacios-Huerta, I., 2004).

Football is widely referred to as "the beautiful game" (Vrooman, J., 2007) and it is rightfully deemed so because it can be played by anyone, anywhere, at any time, and all that is required is a football. A blend of physical, tactical, decision making and technical skills is quintessential to playing this sport, which makes football very complex. The objective is simple: to score more goals than the opposition within a stipulated time, usually 90 minutes.

The simplicity of this sport is what makes it so entertaining and relevant in the current world of economic and cultural globalization (Bairner, A., 2015). The International Federation of Association Football (FIFA) was set up in 1904 with just 7 founding members and has now expanded drastically to 211 associated affiliations (Duval, A. and Heerdt, D., 2020). Events such as the Men's Football World Cup help garner a lot of media attention and help generate tourism and revenue, thereby leading to countries vying to host such events as a matter of prestige and pride (Tomlinson, A. and Young, C. eds., 2006). The FIFA World Cup is the single largest, sports event globally and the 2014 World Cup in Brazil had 3.2 billion people who tuned in to watch the 64 matches (FIFA, 2014a).

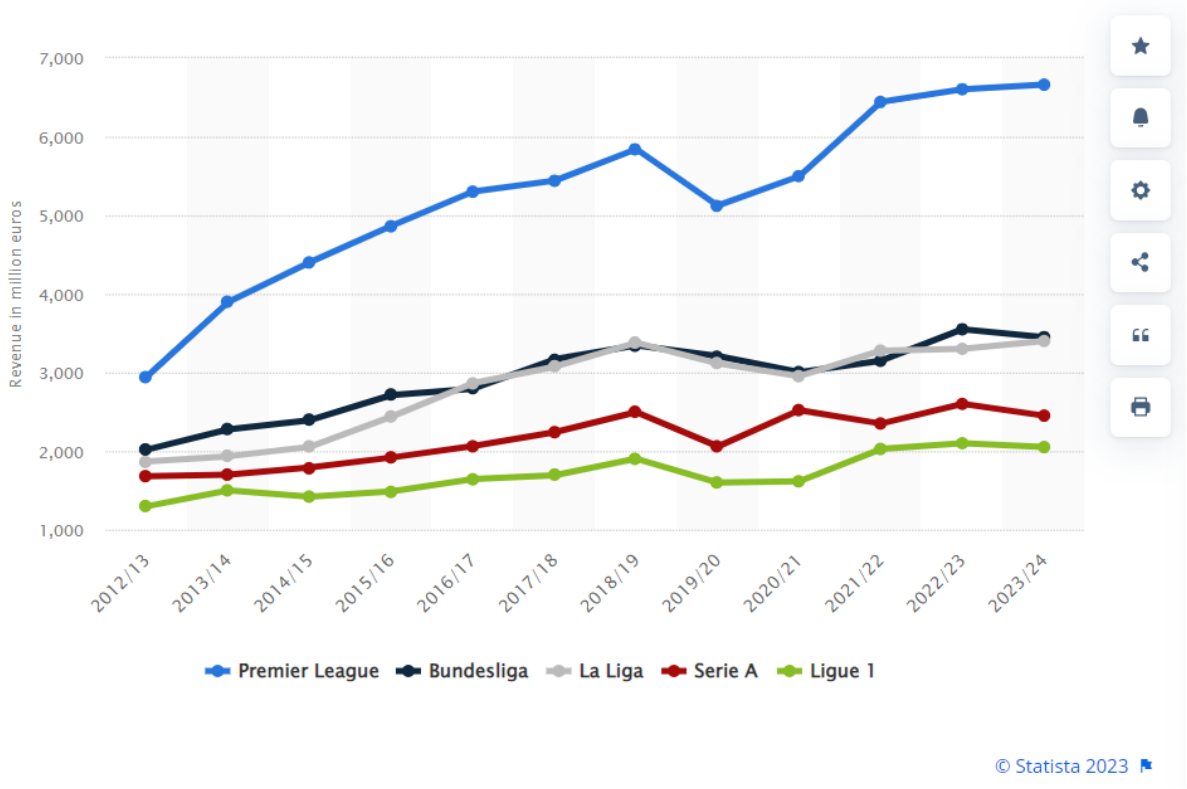
1.2 English Premier League

According to the Union of European Football Associations (UEFA, 2023) the top 5 professional men's football leagues in the world are-

a. English Premier League (England)
b. La Liga (Spain)
c. Serie A (Italy)
d. Bundesliga (Germany)
e. Eredivisie (Netherlands)

Out of these leagues, the English Premier League is renowned for its physicality as well as direct playing style, while the La Liga is characterized by the tiki-taka, possession based play style and the Serie A is distinguished by its defensive prowess (Sarmento, H., Pereira, A., Matos, N., Campaniço, J., Anguera, T.M. and Leitão, J., 2013).

The English Premier League takes place every year and has a total of 20 teams vying for the trophy. The top 4 teams qualify for the UEFA Champions League and the bottom 3 teams get relegated to a lower league called EFL Championship. The English Premier League based in England is the most watched football league in the world (King, A., 2017). The English Premier League is now telecast in 212 countries and has around 4.7 billion people viewing every season, thereby making the English Premier League clubs the most sought after, valuable and richest clubs in the world (Barnes, C., Archer, D.T., Hogg, B., Bush, M. and Bradley, P., 2014).



- **Fig 1** (Statista Research Department. 2023.)

The figure taken from Statista clearly illustrates the revenue generated by the Top 5 leagues from seasons 2012-13 to 2021-22. It clearly depicts the immense revenue disparity between the English Premier League and the other four leagues. This is one of the reasons for selecting this dataset of footballers plying their trade in the English Premier League. Another reason for this study is that Manchester, Liverpool, and London are part of the top 5 most attractive locations for professional footballers to ply their trade (Tobar, F. and Ramshaw, G., 2022).

1.3 Importance of the Research

This dissertation proposes to perform a quantitative analysis of footballer performances in the English Premier League for the 2022-2023 season, examining the performance metrics of footballers based on their position and what are the factors that affect their overall match rating. The study will further leverage the advancements in machine learning algorithms to model and determine whether the ratings given by these machine learning models are similar to those that are predetermined.

Individualistic performances are critical for football players and the clubs they represent to thrive, even if football is a team sport (Wakelam, E., Steuber, V., & Wakelam, J. 2022). As a result of this, new techniques are sought after by analysts, coaches, clubs and scouts in order to recognize emerging talents from different leagues and also homegrown talents at grassroots levels. The revenue gaps prevalent between the traditional big and small clubs is massive (Pifer, N.D., Wang, Y., Scremin, G., Pitts, B.G. and Zhang, J.J., 2018). Due to this revenue gap, the smaller clubs struggle to cope with the big clubs when it comes to spending on buying players as the bigger clubs outbid the smaller clubs easily. The fees for transfers is also getting much higher with every passing season as the market for football players is inflated (He, M., Cachucho, R. and Knobbe, A.J., 2015). For this reason, and with the advent of worldwide technological prowess, football analytics has come into the picture in order for tactical analysis, scouting new talents as well as assessing player performances (Anderson, C. and Sally, D., 2013). This helps ensure that clubs save millions of funds during recruitment and scouting and end up looking for players that the team genuinely needs rather than transferring in players that have performed well but are not required for that particular team or position (Peng, K., Cooke, J., Crockett, A., Shin, D., Foster, A., Rue, J., Williams, R., Valeiras, J., Scherer, W., Tuttle, C., Adams, S., & Rhodes, M. 2018).

Traditionally, football analytics relied on eye test and analysing by observation by watching match compilations. These outdated methods, though necessary, do not tell the entire picture of the statistics put up by the footballer during the course of the match. Furthermore, this study is done since the conventional metrics traditionally used for analysis such as the number of assists, goals scored, clean sheets, etc. are not the only statistics that can help determine whether a player is good or not. By reviewing just these metrics, clubs are missing out on young gems whom they could have purchased for cheap transfer fees as these players go under the radar since the conventional metrics used for football analytics do not deem them as good players (Ruijg, J. and van Ophem, H., 2015). For this reason, this dissertation delves into recognizing the significant contributors that determine the average match rating of a footballer. Since machine learning models are widely used, a comparison between the Multiple Linear Regression Model and the Random Forest Model is done in order to determine which is the better model to determine who is the better footballer based on their respective positions.

1.4 Research Questions

i. What are the key performance indicators (KPIs) that help determine a footballer's overall match rating based on each position?

ii. Which machine learning models can be used to rank these players based on their positions?

iii. Which machine learning model is more optimal when determining the rank of the footballers based on their position?

1.5 Organization

This dissertation delves into the quantitative analysis of footballer performance in the English Premier League for the 2022-23 season using machine learning algorithms. This research is structured into five parts in order to effectively present information in an organized format. The first section broadly explains the background of the sport of football, the popularity of the sport, the major global football leagues, the English Premier League, the need for this research and the research questions of this study. The next section explores what has been done in the field of football analytics, the evolution of football on the whole, the metrics used to analyze footballer performances, position specific requirements in football, machine learning algorithms used, the overall research gap, and the hypotheses for this study. Further on, the methodologies used in this study are introduced and the means of data preprocessing and model selection and evaluation are explored. The analysis of the dataset is explained in-depth in the following section followed by the conclusions made in this study, the limitations, implications and the room for further study.

2. Literature Review

This section explores how football players were scouted, how football analytics has advanced over the years, the ways in which football has evolved, the processes of footballer evaluation, the machine learning algorithms that have been used in this field and the research gap.

2.1 Advent of Football Analytics

In order for teams to compete for trophies, they need to build a team of the best possible eleven players from different nationalities. Acquiring hidden gems often results in teams competing for the same player and puts the smaller clubs at a disadvantage due to the large financial disparity between the big and small clubs (Plumley, D.J., Wilson, R. and Shibli, S., 2017). Since smaller clubs struggle to compete on the exorbitant transfer fees and footballer wages, it led to an increased imbalance competitively (Vöpel, H., 2011). Due to these budget constraints, clubs began to extensively rely on analytics for the purpose of recruiting and scouting (Rein, R. and Memmert, D., 2016). These football clubs rely on analytics to identify talents, optimize the performance of the team, and ensure a sustainable approach to squad building (Lewis, M., 2004).

Data analytics in football began way back in the 1950s when Charles Reep started to manually go through match data to devise long ball tactics based on the rate of loss of ball possession (Reep, C. and Benjamin, B., 1968).

Sophisticated systems were introduced for tracking and collection of data in the major football leagues in the 1990s by companies like Opta and ProZone which allowed for enhanced technical and tactical analysis (James, N., 2006).

The goals scored in a match were predicted by using Poisson Distribution by comparing previous results of club matches (Moroney, M.J., 1956). This statistical model of distribution was also used for determining when a goal would be scored in a game (Dyte, D. and Clarke, S.R., 2000) and also to analyze the goals scored and the location from which they were scored (Carmichael, F., Thomas, D. and Ward, R., 2001).

Another statistical model that was widely used during that time was the Negative Binomial Distribution. It allowed for more variance than Poisson Distribution as can be seen when using this model to study the foul counts in games (Ridder, G., Cramer, J.S. and Hopstaken, P., 1994).It was used to deem how situational variables affect the number of shots taken in a game (Liu, H., Gómez, M.A., Gonçalves, B. and Sampaio, J., 2016) and to model how a team scores goals throughout the season in order to take overdispersion into account (Karlis, D. and Ntzoufras, I., 2003).

In order to devise tactics before the game and to strategize the decision making of the team and the players when it comes to taking a penalty or other in-game decisions, pay-off matrices were introduced (Palacios-Huerta, 2014). To comprehend and analyze how a red card may affect the outcome of a game, (Ridder, G., Cramer, J.S. and Hopstaken, P., 1994) changed the pay-off matrices to come up with different outcomes.

Entropy metrics were utilized to quantify uncertainty and random events in football, which further helped in building the coordination of the team (Liu, H., Gomez, M.Á., Lago-Peñas, C. and Sampaio, J., 2015). The more unpredictable the event is, the higher the entropy. In this regard, Shannon entropy was applied to measure the amount of randomness in events that take place in the match such as tackles, shots or possession (Lucey, P., Oliver, D., Carr, P., Roth, J. and Matthews, I., 2013).

In order to estimate the amount of interactions between defenders and forwards, differential equation models were implemented (Gudmundsson, J. and Horton, M., 2017). To simulate appropriate coverage for the defence and to come up with emergent tactics, this model was used.

2.2 Footballer Performance Metrics

Even though football is a team sport, the individual performances of the footballers collectively help the team ultimately succeed. In order to quantify the performances of a footballer, the following have been implemented.

2.2.1 Conventional Metrics

The importance of a footballer to his team was quantified by using measures of centrality and by utilizing network analysis. Studies have tried to quantify the impact of an individual player on the overall team performance using network analysis and centrality measures. This can help estimate a player's contribution and importance (Cintia, P., Rinzivillo, S. and Pappalardo, L., 2015).

The metric of plus-minus was devised to measure whether the player's presence makes a difference to his club's performance (Grund, T.U., 2012).

Conventionally, the number of goals a player scored or the number of assists provided determines whether a striker is a good player or not (Pollard, R. and Reep, C., 1997).

For midfielders, the players were rated based on their pass completion rates (Jones, P.D., James, N. and Mellalieu, S.D., 2004).

Defenders were rated based on the number of tackles that they made in a game and based on the number of cleansheets accrued (Yiannakos, A. and Armatas, V., 2006).

2.2.2 Advanced Metrics

Using conventional attack minded players this has its disadvantages since this completely depends on team dynamics and the tactics employed by the coaches (Liu, H., Gómez, M.A., Gonçalves, B. and Sampaio, J., 2016). Pass completion rates are not always useful because backward and sideward passes are significantly not as important strategically as compared to a forward pass (Power, P., Ruiz, H., Wei, X. and Lucey, P., 2017).

Sophisticated models such as expected assists (xA) and expected goals (xG) are now used to analyze the performances of a player (Eggels, H., van Elk, R. and Pechenizkiy, M., 2016).

xG is a more accurate measure of how many goals a player potentially scores rather than the actual number of goals he has already scored since this takes the quality of the chance created, the angle and the distance from which the player shoots into account (Fernández, J., Bornn, L. and Cervone, D., 2021).

xA measures the quality of the chance created and quantifies how likely a pass will lead to a goal based on the weight of the pass, angle of the pass and distance from goal (Decroos, T., Bransen, L., Van Haaren, J. and Davis, J., 2019)

The measure of the number of duels won gives a more accurate evaluation as to how effective and valuable a defender or midfielder's contribution is rather than just the number of tackles made (Hughes, M. and Franks, I., 2005).

Heat maps are studied to analyze areas of the pitch where the player is most effective based on his actions on the pitch with and without the ball (Gyarmati, L. and Stanojevic, R., 2016).

The opponent team's playmakers, passing combinations and playstyles are analyzed by visualizing the passing networks (Bialkowski, A., Lucey, P., Carr, P., Yue, Y. and Matthews, I., 2014).

2.3 Footballer Position Analysis

2.3.1 Evolution of Modern Football

Positional analysis and tactical flexibility are paramount to comprehending the ever-evolving demands of modern football. Footballers tend to struggle while moving to the Premier League as it is harder to adapt to than other leagues because the physical intensity of the league is much higher in terms of the distance run overall, accelerations and the number of player sprints in a game (Barnes, C., Archer, D.T., Hogg, B., Bush, M. and Bradley, P., 2014).

Modern football has led to English Premier League teams adopting a more direct style of play and play more long balls as compared to the tiki-taka possession based style of play of other regional football leagues (Carling, C., Le Gall, F. and Dupont, G., 2012). A player's technical attributes such as his first touch and his ability to dribble is tested in the English Premier League due to its intensity (Dellal, A., Chamari, K., Wong, D.P., Ahmaidi, S., Keller, D., Barros, R., Bisciotti, G.N. and Carling, C., 2011).

2.3.2 Position Specific Requirements

Goalkeeper- Goalkeepers are judged by the amount of saves that they make in a game and the number of cleansheets accrued overall. For this reason, they need to have top-notch reflexes, command of the 18 yard penalty area and the awareness to act as sweeper keepers if necessary and to also have technical ability to be able to distribute the ball accurately and start the attack from the back (Castañer, M., Barreira, D., Camerino, O., Anguera, M.T., Canton, A. and Hileno, R., 2016).

Defenders- Defenders are either classified as a Fullback or a Centre Back. The primary metric to evaluate a defender is to determine whether he can defend or not. But, with the evolution of modern football, central defenders are also required to have accurate and long passing range and astute levels of game awareness to handle the high pressing systems being employed by opponent forwards (Lago-Peñas, C. and Dellal, A., 2010) Fullbacks are not just required to be able to defend (Perarnau, M., 2014), but must have incredible fitness levels as they have to provide options in attack by either overlapping or inverting based on team tactics (Taylor, J.B., Mellalieu, S.D. and James, N., 2004).

Midfielders- Central midfielders cover the most distance through high work rates and relentless pressing (Bradley et al., 2013). Central midfielders must dictate tempo, resist opposition pressing, and penetrate compact defences. Excellent passing range and vision is imperative (Liu et al., 2015). Also requires creativity to break down tight defences (Rein et al., 2017). Midfielders face increased defensive duties (Dellal et al., 2011).

Wingers need pace and dribbling skills to create 1v1 situations against fullbacks (Taylor et al., 2017). Also track back to support defense. Midfielders counterpress immediately after losing possession to win back the ball high up the pitch (Fernandez-Navarro et al., 2018). This requires conditioned repeated sprinting.

Forwards- Forwards are not just required to score goals, but also be able to facilitate attacks by linking up with other teammates by being able to hold-up play (Shaw, L. and Glickman, M., 2019) by winning more duels (Lago-Peñas, C. and Dellal, A., 2010). Forwards are also required to be able to exploit the space in behind defenders and occupy goal scoring areas by possessing adept off-ball movement (Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R. and McRobert, A.P., 2016).

2.4 Machine Learning Algorithms

Machine Learning has vastly improved the way that data is analyzed and interpreted in football analytics. Machine learning models help in identifying patterns and trends and help in making predictions on the datasets. These insights aid in player scouting, injury prediction, player transfer strategies (Rein, R. and Memmert, D., 2016) and recovery

programs (Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J. and Medina, D., 2018).

Multiple Linear Regression was used to predict the value of a dependent variable based on one or more independent variables by Sampaio, J., Lago, C., and Drinkwater, E.J. (2010) who predicted the match outcomes by using shots on goal, shoots on target, tackles, and fouls committed. Multiple regression approaches were utilized by researchers to evaluate the economic value of football players based on their in-game data. Carmichael, F., Thomas, D., and Ward, R. (2000) estimated player valuations as a function of goals, games played, position, and transfer cost.

Neural networks has been used for tasks such as evaluating on-ball actions, estimating player market prices, and forecasting match results (Decroos, T., Bransen, L., Van Haaren, J., and Davis, J., 2019).

Deep learning allowed for more intrinsic investigation of team strategies and tactics by collecting situational information from raw tracking data (Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., and Fookes, C., 2016).

2.5 Research Gap

1. Position-Specific Analysis: There is a limited focus on position specific analysis of football players even though comprehensive studies have been done on player performance metrics. Most studies seek to generalize player performance across positions, which may miss the specific contributions and requirements of each position on the field.

2. Comparative Analysis of ML Models: Many studies have used machine learning models to analyze football data, but a complete comparison of different models (such as Multiple Linear Regression and Random Forest) in forecasting player performance, while taking a combination of both the conventional and advanced metrics particularly in the context of the English Premier League, appears to be lacking.

3. Reliability of Popular Rating Systems: Platforms like SofaScore (which provides the match ratings) have grown in popularity in recent years, but there is still a lack of understanding of how their Average Match Ratings are impacted by numerous KPIs. Furthermore, the dependability and validity of these platforms in accurately documenting a player's performance remain unknown.

4. Practical Implications of Advanced Metrics: While sophisticated metrics have been established and are in use, the practical implications of these measures for clubs, particularly in terms of scouting, player development, and tactical decisions, have received little attention.

5. Bridging Traditional and Advanced Metrics: To offer a comprehensive picture of a player's contribution and performance, there is a need to bridge the gap between conventional metrics (such as goals, assists, and tackles) and advanced metrics (such as xG, xA).

2.6 Contribution to the Study and Hypotheses

The study aims to provide a nuanced comprehensive understanding of footballer performance that would offer insights that would aid English Premier League clubs in scouting, strategizing and player development. The hypotheses are as follows:

H1: Different positions in football (Goalkeeper, Forward, Defender, and Midfielder) have distinct KPIs that predominantly determine their Overall Average Match Rating.

H2: Machine Learning algorithms- Multiple Linear Regression and Random Forest, can effectively rank footballers based on their predicted performance ratings derived from their KPIs.

H3: The efficacy of the Machine Learning models in predicting player performance varies based on the player's position:

H3a: For forwards, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.
H3b: For midfielders, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.
H3c: For defenders, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.
H3d: For goalkeepers, Random Forest will provide a more accurate prediction as it has a lesser Mean Absolute Error than that of the Multiple Linear Regression Model.

3. Methodology

This section delves into the types of approaches to research that are prevalent and will focus on the research type implemented in this dissertation. The chapter further goes on to explore the process of how the data was collected and further pre-processed. The reasons for the implementation of machine learning algorithms used as well as the ethical issues are discussed in this chapter as well.

3.1 Research Approach

The different types of research approaches are as follows-

Quantitative/Deductive Approach
This is a type of approach that involves the collection of numerical data. This data is then analyzed using mathematical, statistical or computational methods to test the hypothesis made and to come to conclusions. It is a form of deductive approach where existing theories are researched and hypotheses are then formed and tested (Creswell, J.W. and Creswell, J.D., 2017).

Qualitative/Inductive Approach
This is a type of approach that involves using non-numerical data. This could be in the form of images, interviews, text or video. It focuses on comprehending the process from square one. These open-ended questions are thus framed in order to generate insights and to find patterns and trends (Merriam, S.B. and Tisdell, E.J., 2015). This method of understanding patterns is called the inductive approach.

Mixed Methods
This type of research approach is a combination of quantitative as well as qualitative research approaches. Data is collected from both the qualitative and quantitative sources and the findings of the analysis are integrated to come to a concrete evaluation of the research questions (Creswell, J.W. and Creswell, J.D., 2017).

The research method used in this dissertation is the quantitative / deductive approach. The data collected is numerical data and is analyzed using mathematical and statistical methods to test the hypotheses mentioned in Section 2.6 of this study.

3.2 Data Collection

Secondary data has been used for this study and the data is collected from various sources and is then encapsulated into a single CSV file that is used to finally extrapolate the required conclusions.

The list of the footballers playing in the English Premier League in 2022-23 season is the data sample taken from the official Premier League website: https://www.premierleague.com/

The various attributes required for the player analysis were downloaded from the website: https://footystats.org/download-stats-csv#

This contained a CSV file of the dataset of 673 English Premier League players who belonged to the squad of the 20 teams and had 273 attributes which quantified their performance statistics.

3.3 Data Pre-processing

Before proceeding with the analysis of data, the most fundamental step is that of data preprocessing. In order to make the raw data suitable for analysis, the data needs to be preprocessed since dirty or noisy data hampers the performance of the model (Ji, S., Li, Q., Cao, W., Zhang, P. and Muccini, H., 2020). Null, duplicate or missing values can be handled by either imputing the values or removing them from the dataset (Cheema, J.R., 2014). The presence of outliers is also checked and handled as outliers affect the assumptions and skew the results (Osborne, J.W. and Overbay, A., 2004).

Feature selection is crucial while analyzing the footballer performance because not every attribute of the performance metric may be indicative of the player's overall match rating. In this dissertation, since Multiple Linear Regression is chosen, the attributes are selected after performing Correlation analysis. In order to ensure that the predictors in the model are independent, the multicollinearity is identified and removed (Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J. and Münkemüller, T., 2013). Multicollinearity refers to when two or more predictor variables are highly correlated. In order to check if they are highly correlated or not, various methods are used. The process of checking the correlation matrix for multicollinearity was used in this dissertation. If the correlation coefficient is close to +1 or -1, it indicates high positive or high negative correlation respectively. If two variables are causing multicollinearity, one of them are removed so that the accuracy of the model is not compromised and so that it is still optimal to interpret the individual parameters (O'brien, R.M., 2007).

Since the machine learning algorithms used in this study require numerical data, the categorical data is converted to a numeric format. Some attributes have values that hold greater significance if they are lesser in value and vice versa. In order to handle this, the process of normalization is done in order for the data to be sensitive to the scale of the attributes taken as the input (Scrucca, L. and Serafini, A., 2019). There are various types of normalization techniques such as the min-max normalization which focuses on using the distance metrics to ensure all the values are of a similar scale (Patro, S.G.O.P.A.L. and Sahu, K.K., 2015) or the standard scaling where each attribute is scaled to a standard deviation of 1 and a mean of 0 (Jolliffe, I.T. and Cadima, J., 2016). However, in this study, data type inversion and value inversion have been done for different columns.

3.4 Model Selection

The two machine learning models selected for the analysis in this study are Multiple Linear Regression and Random Forest.

3.4.1 Multiple Linear Regression

Multiple Linear Regression is a machine learning algorithm that extends the principles of simple linear regression which contains just one explanatory variable. It is a statistical model

which helps in predicting the value of a continuous variable based on two or more variables that are explanatory (Field, A., 2013). A linear relationship has to be present between the independent and dependent variables and there must not be multicollinearity between the independent variables (Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021). The Multiple Linear Regression model can be represented by the following equation:

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta nXn + \varepsilon$ (Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021)

Where Y is the dependent variable,
X1…Xn are the independent variables,
β0 is the slope,
β1 to βn are the regression coefficients,
ε is the error.

The method of least squares is used to determine the coefficients and this model uses the R^2 statistic to determine the overall fit of the model (Field, A., 2013).

3.4.2 Random Forest

Random Forest is a machine learning model that combines multiple decision trees while training the model (Breiman, L., 2001). This is done to reduce the variance and the predictions thus made are more accurate than that made by a one decision tree (Breiman, L., 2001). The parameters for this model are the number of trees and the number of variables that are sampled randomly at each split. It is sampled randomly to reduce the variance of the predictions. The Random Forest model outputs the classification or the regression of the individual trees for the given input. Each tree gives a respective classification and the random forest selects the classification with the most votes across all the trees in the forest.

3.5 Model Evaluation

In order to decide which of the two models are best suited for the data analysis, machine learning model evaluation is done by a variety of methods (James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013) such as- Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R-squared).

In this study, the Mean Absolute Error (MAE) method has been employed. Mean Absolute Error is an error metric that calculates the average magnitude of errors between the observed and the predicted values (Hyndman, R.J. and Koehler, A.B., 2006). It treats all errors equally as it assigns equal weights to all errors (Chai, T. and Draxler, R.R., 2014.). It is easy to interpret and represents the average absolute error in the same units as that of the target variable. Mean Absolute Error is robust and not sensitive to outliers, unlike the MSE and RMSE (Willmott, C.J. and Matsuura, K., 2005).

It is mathematically represented by the formula:

$MAE = (1/n) \sum ni=1 |yi - \hat{y}i|$ (Chai, T. and Draxler, R.R., 2014.)

Where:

n = number of observations
yi = actual value
ŷi = predicted value

The Mean Absolute Error is calculated for both the Multiple Linear Regression as well as the Random Forest model based on each position. The model that has the lower Mean Absolute Error for that respective model is the one that must be utilized to predict a footballer's average overall match rating for that particular position (Chai, T. and Draxler, R.R., 2014.).

3.6 Ethical Issues

Secondary data has been used for analysis in this dissertation. Data that is made public has been used and no personal information of any participant has been taken. The sources for the dataset collection have been mentioned in Section 3.2. Therefore, there are no ethical concerns that may arise due to this dissertation.

4. Data Analysis and Findings

This section discusses the analysis of the English Premier League footballer performances for the 2022-23 season. The sampled players have been classified based on their playing position and multiple linear regression and random forest models have been utilized for their performance evaluation. Data was checked to see if it was linear and was also checked for multicollinearity and then chosen for multiple linear regression. The findings of the results of this model helped in identifying the Key Performance Indicators that significantly determine a player's performance based on his playing position. Each of the research questions has been explored and all the stated hypotheses have been proved in this section.

4.1 Identifying Key Performance Indicators

The secondary data was collected of the English Premier League footballers for the 2022-23 season. Only the players of the 20 teams who have had game time on the pitch were considered. This dataset has a list of 506 players from the 20 teams playing in the English Premier League. Out of these players, the footballers have been segregated based on their playing position. This is illustrated below:

Player Position	**Number of Players**
Forwards	112
Midfielders	181
Defenders	174
Goalkeepers	39

Table 1- Number of footballers for each position

The original dataset had 277 variables. This was cleaned to 71 attributes. These remaining attributes were removed since the values were either missing, not relevant or were repetitive. Out of these 71 attributes, many attributes were considered for multiple positions because it is an aspect of football that is required by any footballer who plays in any position on the pitch.

On further examination, it was found that out of the 71 attributes, many were not required to evaluate a footballer's performance based on his position as they did not significantly contribute to his overall average match rating.

In order to analyze these attributes, they were converted to numeric. Even though the CSV of the dataset was manually cleaned, a check was done in R Programming to check for the presence of missing values. The results of this illustrated that there were no missing values in the dataset as illustrated in the image below:

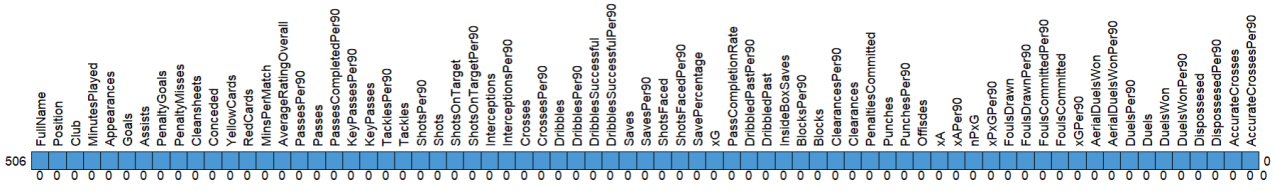


Fig 2- Check for missing values

The dataset was also checked for the presence of outliers. The following variables were shown to have no outliers- MinutesPlayed, Appearances, Cleansheets, Conceded, MinsPerMatch and the outliers present for the other variables are values that are valid and must be considered.

4.2 Descriptive Statistics

Based on the existing literature that explored the footballer performances based on player positions, descriptive statistics of the footballers was derived.

While analyzing the performances of Forwards in modern day football, a Forward is evaluated not just by how many goals he scores, but by his overall game play and by what he brings to the team. In this regard, the forwards who have played more than 500 minutes in the 2022-23 season were filtered. The number of duels a forward contested is analyzed along with the number of shots a forward has taken. These two attributes were considered as it helps to determine which forwards press hard and compete for duels to win the ball and not just rely on assists being provided to them by other teammates and the number of shots taken determines how often a forward decides that he has a goalscoring opportunity due to which he takes a shot. The Fig 3 below illustrates this-

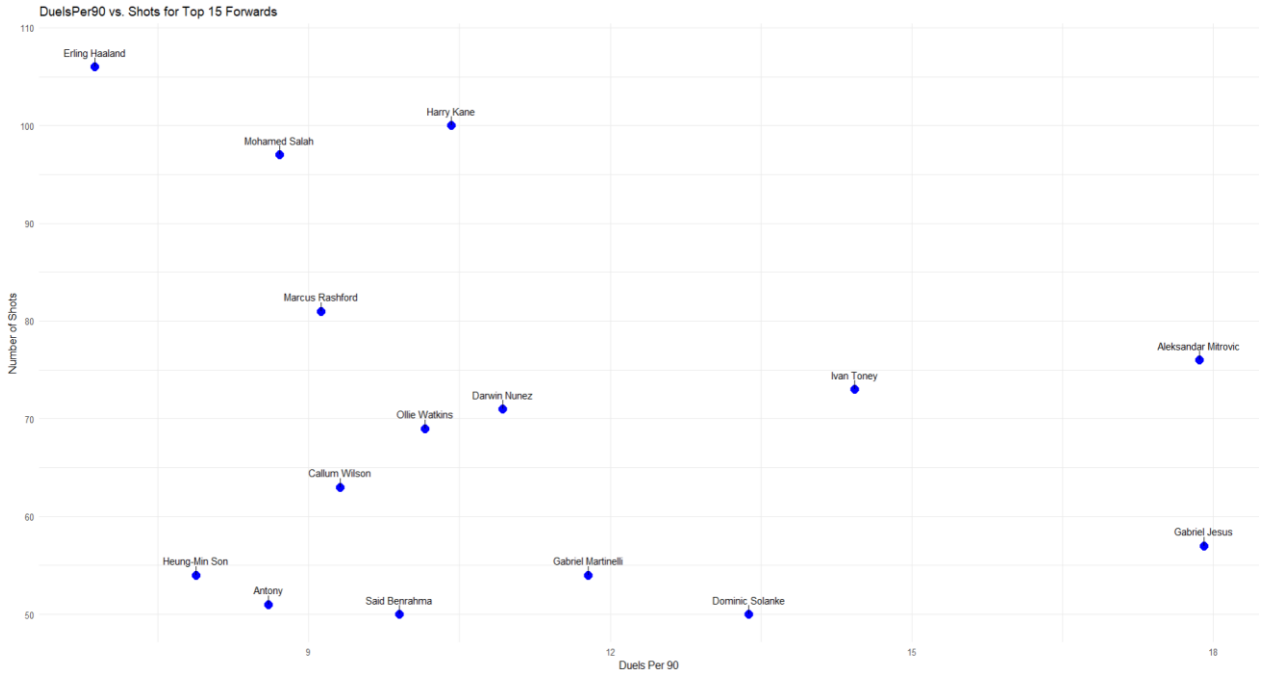


Fig 3

It is clearly seen that in Fig 3, forwards like Erling Haaland, take a lot of shots, but do not compete for a lot of duels, while forwards like Mitrovic and Ivan Toney have a more all-round game where they help their team by competing for a lot of duels and also take a lot of shots as well.

While analyzing the performances of midfielders, a midfielder is not just required to have accurate passing ability. With the evolving tactics and intensity of the modern game, midfielders are required to create a lot of chances as well as provide crosses for the team to

switch play as well to bring other players into the game. The Fig 4 below, clearly illustrates this:

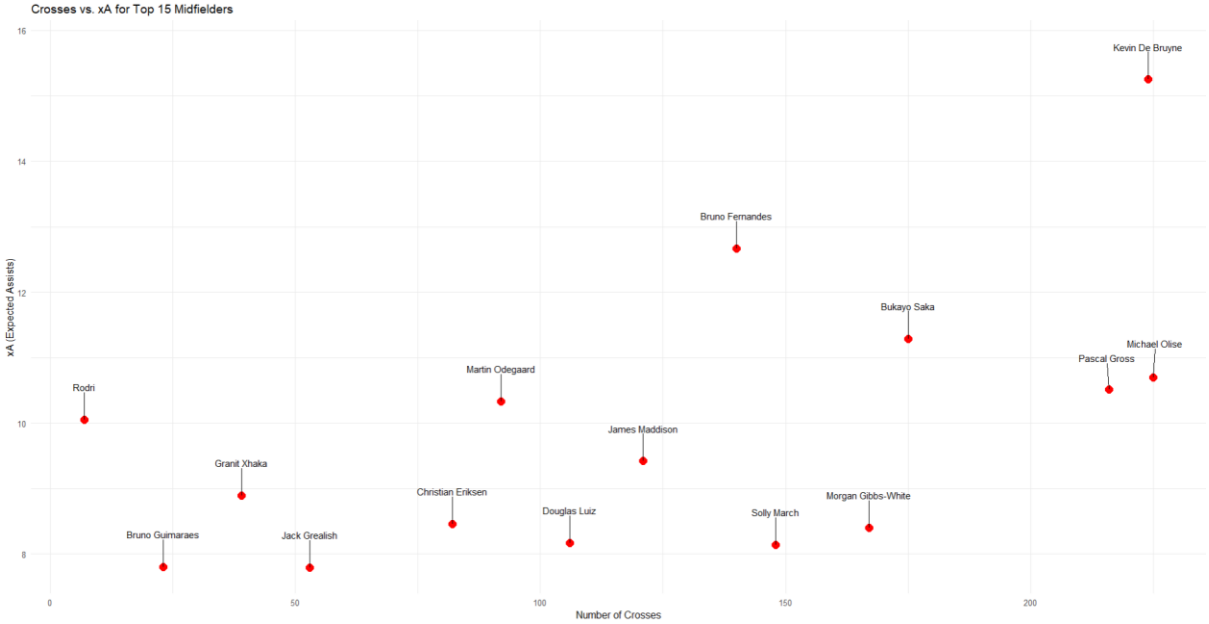


Fig 4

The above figure depicts that Kevin De Bruyne is in a league of his own as a midfielder as he has created the most chances which is measured by xA- which refers to the expected assists, and has provided the most crosses as well.

While analyzing the performances of defenders, conventionally only the number of tackles a player made was taken into consideration. But, the demands of the game require a defender to be able to soak up the pressure of the opponents by having possession of the ball and by recycling the ball by passing. Therefore, Fig 5 analyzes the top 10 defenders with the most tackles as well as passes.

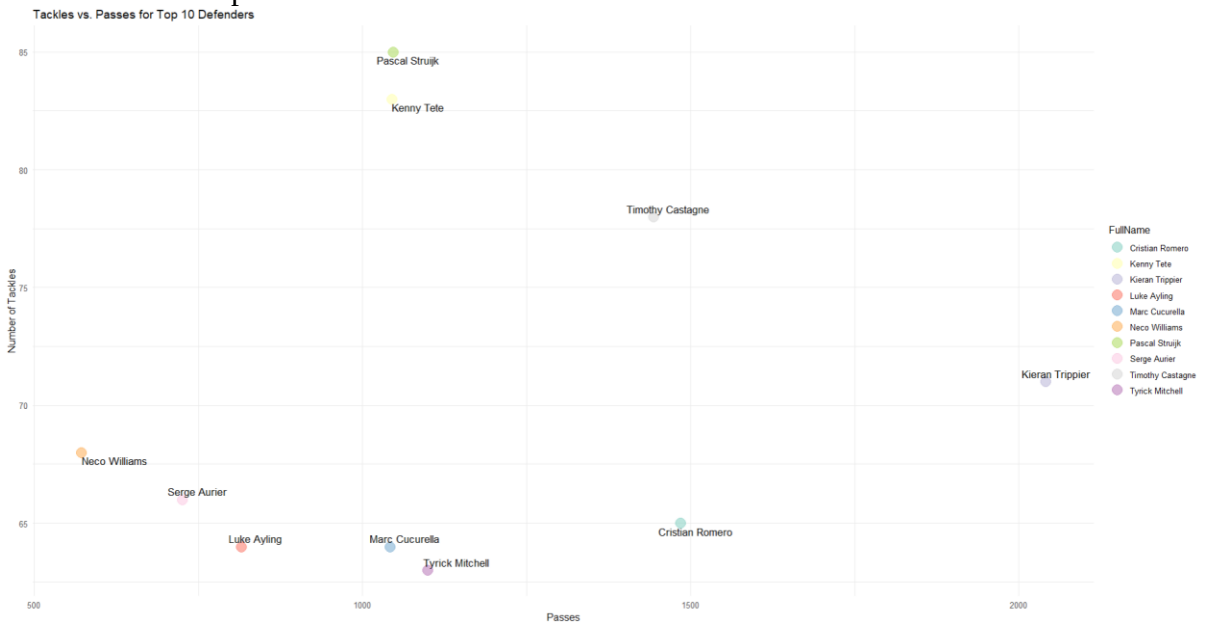


Fig 5

It is evident that Pascal Strujik stands out in terms of the number of tackles made, but has not made any passes. This could mean that he is not very comfortable on the ball or that he loses possession often or that he plays for a team that does not have a lot of possession. But, Timothy Castagne is someone who is a more well-rounded defender as he has made a lot of passes and has made the third highest number of tackles.

While analyzing goalkeepers, the amount of saves made is what determines how effective he is as a shot-stopper. But, since goalkeepers are also required to start attacks since most teams prefer playing out from the back, the pass completion rate of goalkeepers is also considered in Fig 6.

Fig 6

It is clearly seen that David Raya has the highest save percentage, Alisson Becker has the most accurate pass completion rate, but Kepa Arrizabalaga has a very good pass completion rate as well as save percentage and can be considered as a modern day goalkeeper who excels in both attributes.

For further analysis of the attributes, the variable names were converted to numeric. There are 11 attributes whose values are inversed because the lesser the value of these attributes, the better it is for the player's overall performance. Following this, the attributes were normalized to a single scale so that the dataset could be analyzed further.

4.3 Correlation Analysis

Correlation analysis is then done for the 4 subsets of data based on their positions. This is done in order to measure the strength of the relationship between the attributes (Schober, P., Boer, C. and Schwarte, L.A., 2018)
The correlation analysis was performed for the attributes of footballers based on position. This multicollinearity check was done before utilizing these attributes for multiple linear regression (Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q.P. and Lillard Jr, J.W., 2014).

Fig 7 below illustrates the correlation check that was done for 18 attributes for Forwards.

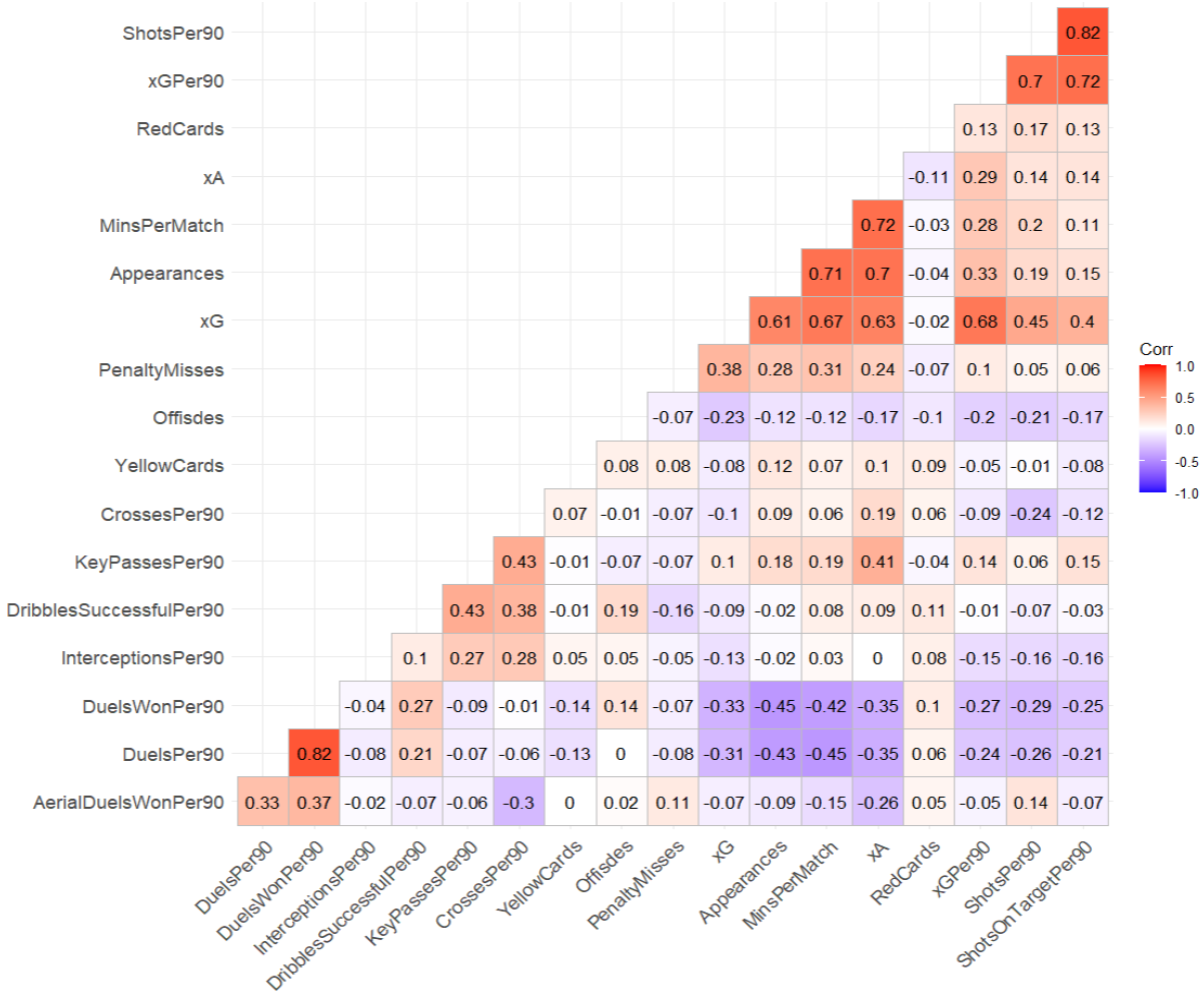


Fig 7

The values that are higher than 0.8 are said to be highly correlated with each other. Therefore, only ShotsOnTargetPer90 and DuelsWonPer90 are the attributes that must be removed, but since they are very close to the value of 0.8, they are considered in this case.

The same analysis was done for midfielders as well by passing 22 attributes as seen in Fig 8.

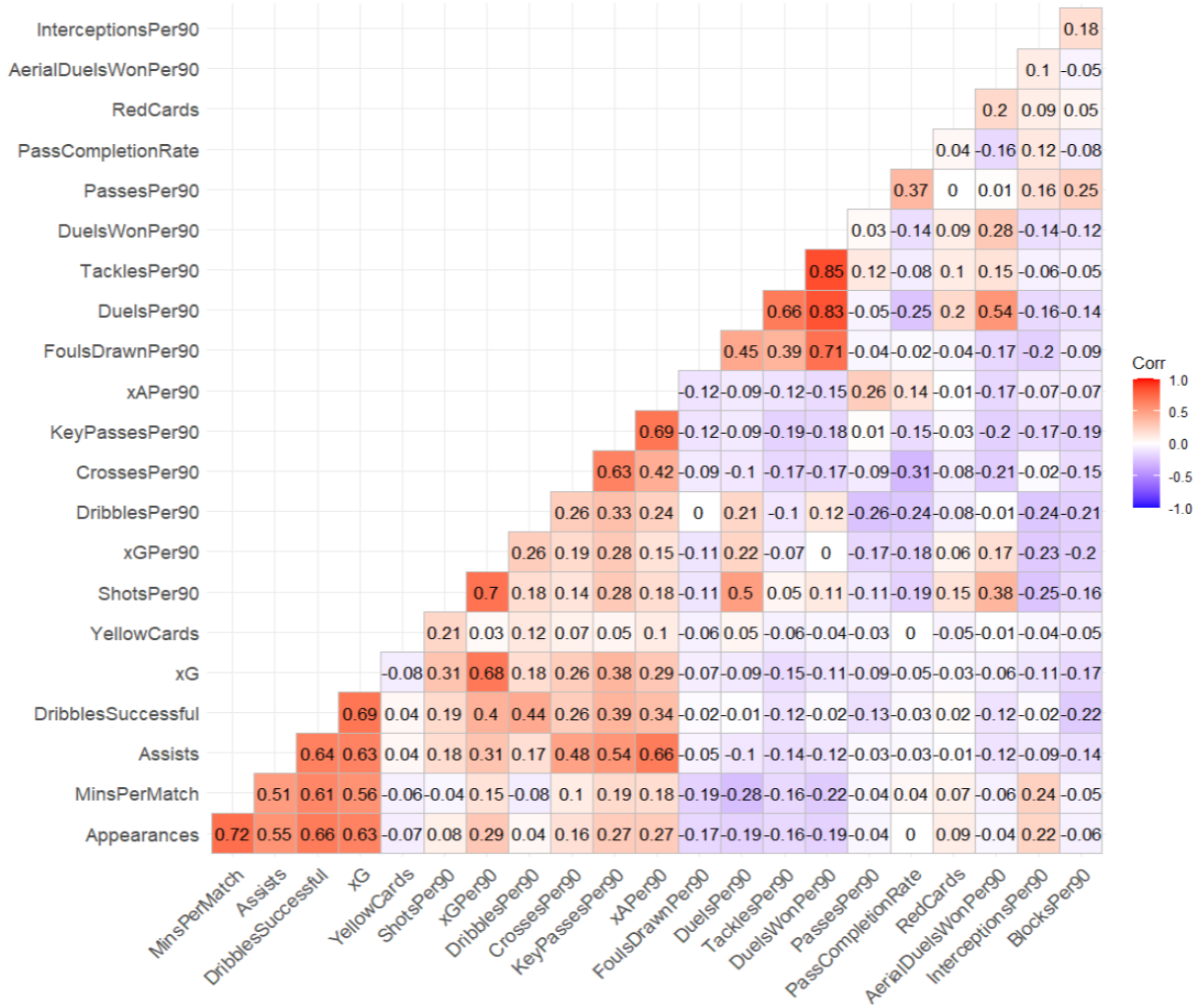


Fig 8

From Fig 8, since the value of DuelsWonPer90 exceeds 0.8, it is removed and not passed for the multiple linear regression model.

While analyzing the attributes for defenders, 25 attributes were considered as illustrated in Fig 9-

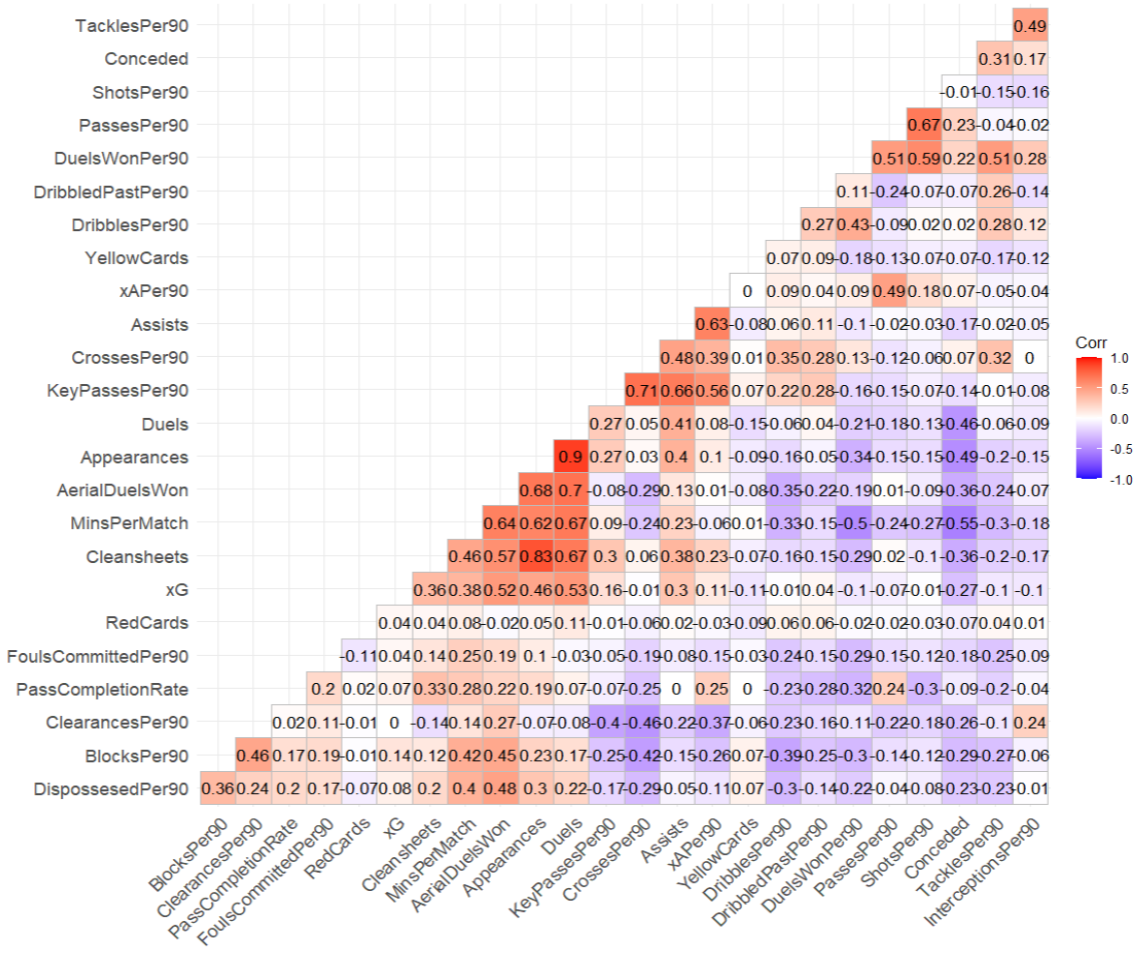


Fig 9

As seen in Fig 9, the variables- Appearances and Duels exceed the threshold of 0.8 and since they are highly correlated, they are removed before passing the other attributes for multiple linear regression.

While checking for multicollinearity for goalkeepers, 16 attributes were considered as seen in Fig 10 below-

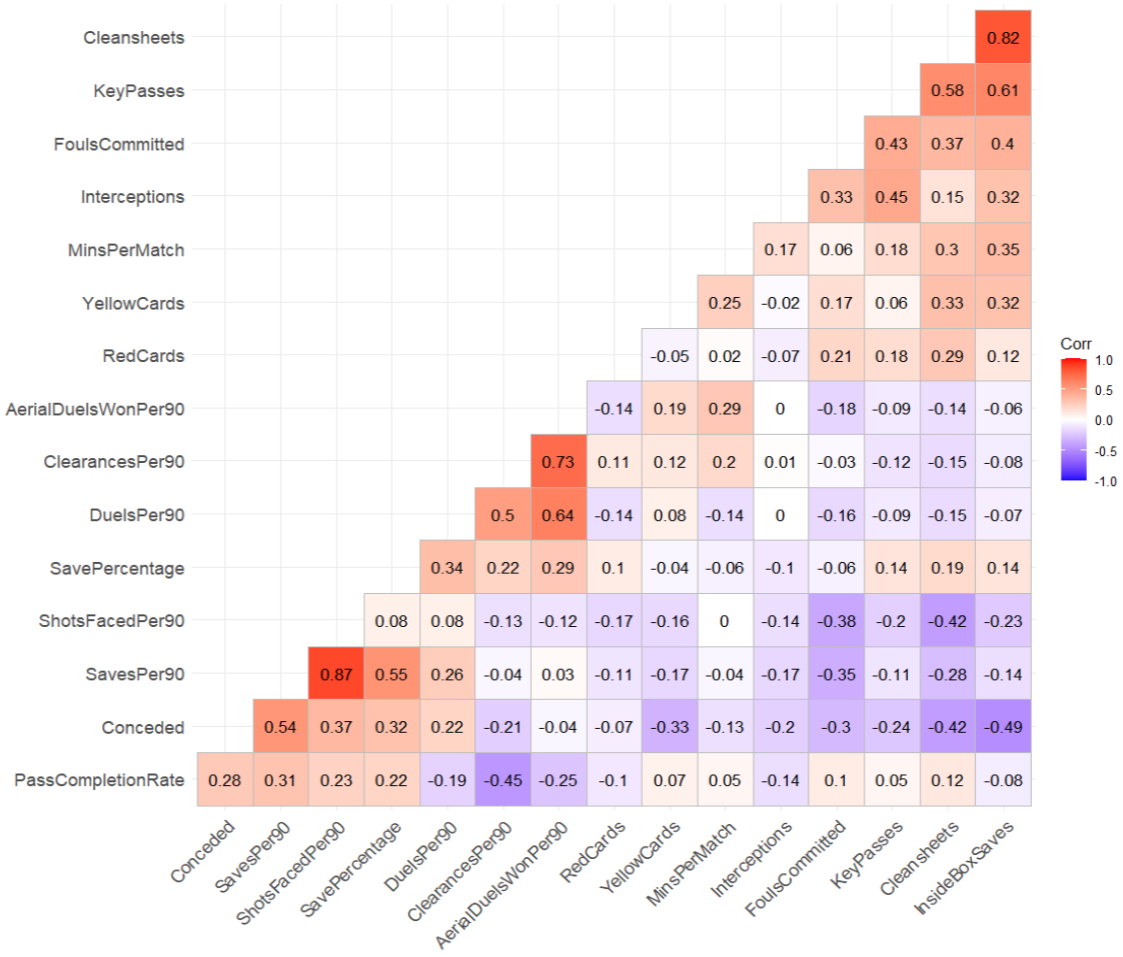


Fig 10

As the value of ShotsFacedPer90 is much more than 0.8, it is removed since it is highly correlated and the other attributes are taken for multiple linear regression.

4.4 Identifying KPIs- Multiple Linear Regression

In order to identify the attributes that contribute to a player's overall match rating, a multiple linear regression model is used. The attributes that can be used for multiple linear regression must be checked for the presence of outliers, checked for linearity of the dependent and independent variables and must also be checked for multicollinearity (Williams, M.N., Grajales, C.A.G. and Kurkiewicz, D., 2013).

The AverageRatingOverall variable is taken as the dependent or response variable for the multiple linear regression model for each position and the other attributes taken for each position are the independent variables or predictors.

After running the 18 attributes through a multiple linear regression model for Forwards, the results are seen in Fig 11 below-

```
> summary(MLRFor)

Call:
lm(formula = AverageRatingOverall ~ Appearances + PenaltyMisses +
    YellowCards + RedCards + MinsPerMatch + KeyPassesPer90 +
    ShotsPer90 + ShotsOnTargetPer90 + InterceptionsPer90 + CrossesPer90 +
    DribblesSuccessfulPer90 + xG + Offisdes + xA + xGPer90 +
    AerialDuelsWonPer90 + DuelsPer90 + DuelsWonPer90, data = InputNorFor)

Residuals:
     Min       1Q   Median       3Q      Max
-2.80394 -0.30665  0.07756  0.40506  2.55486

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.847e-16  6.736e-02   0.000 1.000000
Appearances             -5.446e-03  1.151e-01  -0.047 0.962354
PenaltyMisses           -5.532e-02  7.940e-02  -0.697 0.487685
YellowCards             -2.929e-03  7.351e-02  -0.040 0.968295
RedCards                -1.216e-01  7.332e-02  -1.658 0.100652
MinsPerMatch             5.000e-01  1.268e-01   3.945 0.000155 ***
KeyPassesPer90          -7.057e-02  9.883e-02  -0.714 0.476975
ShotsPer90               2.697e-01  1.446e-01   1.864 0.065423 .
ShotsOnTargetPer90       6.488e-02  1.404e-01   0.462 0.645082
InterceptionsPer90       1.773e-02  7.513e-02   0.236 0.813952
CrossesPer90             6.526e-02  8.764e-02   0.745 0.458330
DribblesSuccessfulPer90  1.753e-01  9.142e-02   1.917 0.058295 .
xG                       3.740e-02  1.527e-01   0.245 0.807061
Offisdes                 9.675e-02  7.561e-02   1.280 0.203863
xA                       5.055e-02  1.340e-01   0.377 0.706850
xGPer90                 -7.300e-02  1.361e-01  -0.536 0.592990
AerialDuelsWonPer90     -5.426e-02  9.296e-02  -0.584 0.560820
DuelsPer90               2.422e-01  1.258e-01   1.926 0.057211 .
DuelsWonPer90           -3.862e-01  1.332e-01  -2.899 0.004667 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7129 on 93 degrees of freedom
Multiple R-squared:  0.5742,    Adjusted R-squared:  0.4918
F-statistic: 6.967 on 18 and 93 DF,  p-value: 8.706e-11
```

Fig 11

The attributes with the least p-values are the values that most significantly affect the dependent variable- AverageRatingOverall.

The number of stars (*) represents how significantly it contributes to the dependent variable. In the case of Forwards, the amount of minutes played per match is the variable that affects a Forward’s Average overall match rating the most as it is represented by three stars (***). This is illustrated clearly in Table 2 below-

Serial No.	Independent Variable	Description
1	MinsPerMatch	Number of minutes played in a match on average
2	DuelsWonPer90	Number of Duels won every 90 minutes
3	DuelsPer90	Number of Duels contested every 90 minutes
4	DribblesSuccesfulPer90	Number of successful dribbles in 90 minutes
5	ShotsPer90	Number of shots taken every 90 minutes

Table 2

After running the 21 attributes through a multiple linear regression model for Midfielders, the results are seen in Fig 12 below-

```
> summary(MLRMid)

Call:
lm(formula = AverageRatingOverall ~ Appearances + Assists + YellowCards +
    RedCards + MinsPerMatch + PassesPer90 + KeyPassesPer90 +
    TacklesPer90 + ShotsPer90 + InterceptionsPer90 + CrossesPer90 +
    DribblesPer90 + DribblesSuccessful + xG + PassCompletionRate +
    BlocksPer90 + xAPer90 + FoulsDrawnPer90 + xGPer90 + AerialDuelsWonPer90 +
    DuelsPer90, data = InputNorMid)

Residuals:
     Min       1Q   Median       3Q      Max
-2.72008 -0.21063  0.03626  0.34692  2.12462

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.806e-16  4.912e-02   0.000   1.0000
Appearances          1.033e-02  8.426e-02   0.123   0.9025
Assists              1.048e-01  1.026e-01   1.022   0.3084
YellowCards         -8.588e-02  5.349e-02  -1.606   0.1104
RedCards             1.034e-03  5.333e-02   0.019   0.9846
MinsPerMatch         5.155e-01  8.712e-02   5.917 1.94e-08 ***
PassesPer90         -1.138e-01  6.312e-02  -1.803   0.0733 .
KeyPassesPer90       1.133e-01  9.005e-02   1.258   0.2103
TacklesPer90         1.290e-01  9.392e-02   1.374   0.1714
ShotsPer90          -7.108e-02  1.076e-01  -0.661   0.5097
InterceptionsPer90  -1.157e-02  6.137e-02  -0.189   0.8507
CrossesPer90        -2.236e-02  7.495e-02  -0.298   0.7658
DribblesPer90        9.853e-02  7.904e-02   1.247   0.2144
DribblesSuccessful   1.914e-02  1.030e-01   0.186   0.8529
xG                  -1.443e-01  1.148e-01  -1.257   0.2107
PassCompletionRate   1.618e-02  6.619e-02   0.244   0.8072
BlocksPer90          7.341e-02  5.748e-02   1.277   0.2034
xAPer90             -7.550e-02  9.804e-02  -0.770   0.4424
FoulsDrawnPer90      9.309e-03  8.797e-02   0.106   0.9159
xGPer90              2.093e-01  1.044e-01   2.005   0.0467 *
AerialDuelsWonPer90 -1.362e-01  9.399e-02  -1.449   0.1492
DuelsPer90          -3.957e-01  1.741e-01  -2.272   0.0244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6608 on 159 degrees of freedom
Multiple R-squared:  0.6143,    Adjusted R-squared:  0.5634
F-statistic: 12.06 on 21 and 159 DF,  p-value: < 2.2e-16
```

Fig 12

In the case of Midfielders, the number of minutes played is what significantly affects the player’s average overall match rating the most as is denoted by the three stars (***). The other key independent variables that affect a midfielder’s average overall match ratings are illustrated in Table 3 below-

Serial No.	Independent Variable	Description
1	MinsPerMatch	Number of minutes played in a match on average
2	DuelsPer90	Number of Duels contested every 90 minutes
3	xGPer90	Amount of expected goals a player should have scored every 90 minutes.
4	PassesPer90	Number of passes made every 90 minutes
5	YellowCards	Number of yellow cards received every 90 minutes
6	AerialDuelsWonPer90	Number of aerial duels won every 90 minutes

Table 3

After running the 23 attributes through a multiple linear regression model for Defenders, the results are seen in Fig 13 below-

```
> summary(MLRDef)

Call:
lm(formula = AverageRatingOverall ~ Assists + Cleansheets + Conceded +
    YellowCards + RedCards + MinsPerMatch + PassesPer90 + KeyPassesPer90 +
    TacklesPer90 + ShotsPer90 + InterceptionsPer90 + CrossesPer90 +
    DribblesPer90 + PassCompletionRate + DribbledPastPer90 +
    BlocksPer90 + ClearancesPer90 + xAPer90 + FoulsCommittedPer90 +
    xG + AerialDuelsWon + DuelsWonPer90 + DispossesedPer90, data = InputNorDef)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47800 -0.22611  0.00441  0.21839  1.89912

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4.723e-16  4.481e-02   0.000 1.000000
Assists            -1.090e-01  8.035e-02  -1.356 0.177008
Cleansheets         2.000e-01  7.476e-02   2.675 0.008295 **
Conceded            3.484e-02  7.189e-02   0.485 0.628673
YellowCards         1.527e-01  4.931e-02   3.096 0.002339 **
RedCards            4.488e-02  4.687e-02   0.958 0.339765
MinsPerMatch        7.775e-01  8.780e-02   8.856 2.19e-15 ***
PassesPer90        -3.663e-02  1.262e-01  -0.290 0.772114
KeyPassesPer90      2.273e-02  9.606e-02   0.237 0.813248
TacklesPer90        3.071e-01  8.722e-02   3.521 0.000570 ***
ShotsPer90         -4.861e-02  1.135e-01  -0.428 0.669137
InterceptionsPer90  5.713e-02  6.224e-02   0.918 0.360153
CrossesPer90        1.784e-01  8.838e-02   2.019 0.045289 *
DribblesPer90      -1.167e-01  7.200e-02  -1.621 0.107147
PassCompletionRate  1.987e-03  7.669e-02   0.026 0.979364
DribbledPastPer90   6.220e-03  5.972e-02   0.104 0.917185
BlocksPer90        -9.595e-03  6.020e-02  -0.159 0.873570
ClearancesPer90     2.315e-01  6.898e-02   3.357 0.000999 ***
xAPer90             1.303e-01  1.010e-01   1.289 0.199333
FoulsCommittedPer90 6.162e-02  5.063e-02   1.217 0.225486
xG                  1.347e-01  5.988e-02   2.249 0.025950 *
AerialDuelsWon     -2.181e-01  9.670e-02  -2.255 0.025574 *
DuelsWonPer90       1.505e-02  1.291e-01   0.117 0.907317
DispossesedPer90    4.140e-02  5.578e-02   0.742 0.459116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.591 on 150 degrees of freedom
Multiple R-squared:  0.6971,    Adjusted R-squared:  0.6507
F-statistic: 15.01 on 23 and 150 DF,  p-value: < 2.2e-16
```

Fig 13

In the case of Defenders, the number of minutes a defender plays per match is what significantly affects his average overall rating the most, while the number of tackles made every 90 minutes and the number of clearances made every 90 minutes are variables that significantly contribute to his average overall match rating as well. The attributes are illustrated in Table 4 below-

Serial No.	Independent Variable	Description
1	MinsPerMatch	Number of minutes played in a match on average
2	TacklesPer90	Number of tackles made every 90 minutes
3	ClearancesPer90	Number of clearances made every 90 minutes
4	YellowCards	Number of yellow cards received every 90 minutes
5	Cleansheets	Number of cleansheets accrued in the season
6	AerialDuelsWon	Number of aerial duels won
7	xG	Amount of expected goals a player should have scored
8	CrossesPer90	Number of crosses made every 90 minutes

Table 4

After running the 15 attributes through a multiple linear regression model for Goalkeepers, the results are seen in Fig 14 below-

```
> summary(MLRGK)

Call:
lm(formula = AverageRatingOverall ~ Cleansheets + Conceded +
    YellowCards + RedCards + MinsPerMatch + KeyPasses + Interceptions +
    SavesPer90 + SavePercentage + PassCompletionRate + InsideBoxSaves +
    ClearancesPer90 + FoulsCommitted + AerialDuelsWonPer90 +
    DuelsPer90, data = InputNorGK)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24385 -0.09919  0.00931  0.11881  0.86290

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.900e-16  7.443e-02   0.000 1.000000
Cleansheets        -1.807e-01  2.108e-01  -0.857 0.400216
Conceded            2.718e-01  1.389e-01   1.957 0.062613 .
YellowCards         1.628e-02  9.039e-02   0.180 0.858635
RedCards           -2.303e-02  9.218e-02  -0.250 0.804928
MinsPerMatch        6.428e-01  1.295e-01   4.964  5.1e-05 ***
KeyPasses          -6.573e-03  1.120e-01  -0.059 0.953709
Interceptions      -1.146e-03  9.485e-02  -0.012 0.990465
SavesPer90         -1.946e-01  1.522e-01  -1.279 0.213824
SavePercentage      6.851e-01  1.485e-01   4.613 0.000122 ***
PassCompletionRate  6.000e-02  1.190e-01   0.504 0.618928
InsideBoxSaves      2.473e-01  2.005e-01   1.233 0.229895
ClearancesPer90     8.834e-02  1.554e-01   0.569 0.575180
FoulsCommitted     -1.192e-01  1.073e-01  -1.111 0.278120
AerialDuelsWonPer90 -6.023e-02  1.735e-01  -0.347 0.731665
DuelsPer90         -1.889e-02  1.433e-01  -0.132 0.896296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4648 on 23 degrees of freedom
Multiple R-squared:  0.8692,    Adjusted R-squared:  0.784
F-statistic: 10.19 on 15 and 23 DF,  p-value: 7.523e-07
```

Fig 14

In the case of Goalkeepers, the number of minutes played followed by the Save Percentage are the most significant contributors to a Goalkeeper’s average overall match rating as seen by the three stars (***). The other key attributes are illustrated in Table 5 below-

Serial No.	Independent Variable	Description
1	MinsPerMatch	Number of minutes played in a match on average
2	SavePercentage	Percentage of saves made
3	Conceded	Number of goals conceded
4	SavesPer90	Number of saves made every 90 minutes
5	InsideBoxSaves	Number of saves made for shots from inside the box

Table 5

The attributes that significantly contribute to a player's Average Overall Match Rating for every given position are identified by the processes above.

4.5 Model Evaluation

4.5.1 Multiple Linear Regression

Multiple Linear Regression Models are run for each of the positions once again. During this run of the Multiple Linear Regression Model, only the above-identified significant contributors to a player's Average Overall Match Rating based on his position are taken.

Multiple Linear Regression- Forwards: This results of this model is clearly illustrated in Fig 15 below-

```
> summary(MLRForKPI)

Call:
lm(formula = AverageRatingOverall ~ MinsPerMatch + ShotsPer90 +
    DribblesSuccessfulPer90 + DuelsPer90 + DuelsWonPer90, data = InputNorFor)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7091 -0.3142  0.0537  0.3522  2.8389

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.863e-16  6.566e-02   0.000 1.000000
MinsPerMatch             4.976e-01  7.663e-02   6.493 2.78e-09 ***
ShotsPer90               2.052e-01  6.922e-02   2.965 0.003745 **
DribblesSuccessfulPer90  2.000e-01  7.031e-02   2.845 0.005331 **
DuelsPer90               1.906e-01  1.167e-01   1.633 0.105356
DuelsWonPer90           -4.050e-01  1.179e-01  -3.436 0.000845 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6949 on 106 degrees of freedom
Multiple R-squared:  0.5388,    Adjusted R-squared:  0.5171
F-statistic: 24.77 on 5 and 106 DF,  p-value: < 2.2e-16
```

Fig 15

The Adjusted R-Squared value of 0.52 is achieved for this model. This value is important as it indicates the goodness of fit of the model, and the higher the value, the better is the fit (Draper, N.R. and Smith, H., 1998).

The Mean Absolute Error is also calculated for this Multiple Linear Regression Model as this will be used to determine which model is better suited to predict a Forward's Average Overall

Match Rating. The results of the Mean Absolute Error for Multiple Linear Regression Model for Forwards is depicted in Fig 16 below-

```
> MAEForKPI
[1] 0.4598583
```

Fig 16

Multiple Linear Regression- Midfielders: This results of this model is clearly illustrated in Fig 17 below-

```
> summary(MLRMidKPI)

Call:
lm(formula = AverageRatingOverall ~ MinsPerMatch + PassesPer90 +
    xGPer90 + DuelsPer90 + YellowCards + AerialDuelsWonPer90,
    data = InputNorMid)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56436 -0.22796  0.03801  0.35012  2.31666

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.761e-16  4.835e-02   0.000 1.000000
MinsPerMatch       5.299e-01  5.218e-02  10.155  < 2e-16 ***
PassesPer90       -1.144e-01  4.933e-02  -2.319 0.021535 *
xGPer90            1.031e-01  5.179e-02   1.991 0.048080 *
DuelsPer90        -2.600e-01  6.206e-02  -4.189 4.45e-05 ***
YellowCards       -9.333e-02  4.869e-02  -1.917 0.056913 .
AerialDuelsWonPer90 -2.217e-01  5.838e-02  -3.797 0.000202 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6504 on 174 degrees of freedom
Multiple R-squared:  0.5911,    Adjusted R-squared:  0.577
F-statistic: 41.91 on 6 and 174 DF,  p-value: < 2.2e-16
```

Fig 17

The Adjusted R-Squared value of 0.58 is achieved for this model.
The results of the Mean Absolute Error for Multiple Linear Regression Model for Midfielders is depicted in Fig 18 below-

```
> MAEMidKPI
[1] 0.4421791
```

Fig 18

Multiple Linear Regression- Defenders: This results of this model is clearly illustrated in Fig 19 below-

```
> summary(MLRDefKPI)

Call:
lm(formula = AverageRatingOverall ~ Cleansheets + YellowCards +
    MinsPerMatch + TacklesPer90 + CrossesPer90 + ClearancesPer90 +
    xG + AerialDuelsWon, data = InputNorDef)

Residuals:
     Min       1Q   Median       3Q      Max
-2.86566 -0.24474 -0.00756  0.26850  2.34706

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.933e-16  4.474e-02   0.000  1.00000
Cleansheets      1.940e-01  6.114e-02   3.172  0.00180 **
YellowCards      1.497e-01  4.636e-02   3.230  0.00150 **
MinsPerMatch     7.872e-01  6.055e-02  13.001  < 2e-16 ***
TacklesPer90     3.233e-01  5.024e-02   6.436 1.28e-09 ***
CrossesPer90     1.555e-01  5.514e-02   2.820  0.00539 **
ClearancesPer90  2.305e-01  5.407e-02   4.262 3.39e-05 ***
xG               8.502e-02  5.407e-02   1.572  0.11777
AerialDuelsWon  -1.613e-01  7.513e-02  -2.147  0.03322 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5902 on 165 degrees of freedom
Multiple R-squared:  0.6677,    Adjusted R-squared:  0.6516
F-statistic: 41.45 on 8 and 165 DF,  p-value: < 2.2e-16
```

Fig 19

The Adjusted R-Squared value of 0.65 is achieved for this model.
The results of the Mean Absolute Error for Multiple Linear Regression Model for Defenders is depicted in Fig 20 below-

```
> MAEDefKPI
[1] 0.3849133
```

Fig 20

Multiple Linear Regression- Goalkeepers: This results of this model is clearly illustrated in Fig 21 below-

```
> summary(MLRGkKPI)

Call:
lm(formula = AverageRatingOverall ~ Conceded + SavePercentage +
    SavesPer90 + InsideBoxSaves + MinsPerMatch, data = InputNorGK)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27048 -0.14175  0.00956  0.11575  0.87440

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.930e-16  6.548e-02   0.000   1.0000
Conceded        2.248e-01  9.302e-02   2.417   0.0213 *
SavePercentage  6.392e-01  8.442e-02   7.572 1.03e-08 ***
SavesPer90     -7.123e-02  8.967e-02  -0.794   0.4327
InsideBoxSaves  3.802e-02  8.720e-02   0.436   0.6657
MinsPerMatch    6.593e-01  7.189e-02   9.171 1.35e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4089 on 33 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:  0.8328
F-statistic: 38.85 on 5 and 33 DF,  p-value: 6.729e-13
```

Fig 21

The Adjusted R-Squared value of 0.83 is achieved for this model.
The results of the Mean Absolute Error for Multiple Linear Regression Model for Goalkeepers is depicted in Fig 22 below-

```
> MAEGkKPI
[1] 0.245618
```

Fig 22

4.5.2 Random Forest

Random Forest Models were run for each of the positions while taking only the significant attributes that affect a footballer's average overall match rating based on his position. The Models were run and the Mean Absolute Errors were calculated for each Random Forest model for the respective position and the following results were obtained.

Random Forest- MAE (Forwards): The Mean Absolute Error found for the Random Forest Model of Forwards is seen in Fig 23 below-

```
> set.seed(12345)
> TrainForCountRF <- sample(nrow(InputNorFor), 0.8 * nrow(InputNorFor))
> TrainForRF <- InputNorFor[, c("AverageRatingOverall","MinsPerMatch","ShotsPer90","DribblesSuccessfulPer9
0","DuelsPer90","DuelsWonPer90")][TrainForCountRF, ]
> TestForRF <- InputNorFor[, c("AverageRatingOverall","MinsPerMatch","ShotsPer90","DribblesSuccessfulPer9
0","DuelsPer90","DuelsWonPer90")][-TrainForCountRF, ]
> #Training
> RF_For <- randomForest(AverageRatingOverall ~ ., data = TrainForRF)
> #Rank Prediction
> TestForRF$PredictedRanking <- predict(RF_For, newdata = TestForRF)
> InputNorFor$PredictedRankingRF <- predict(RF_For, newdata = InputNorFor)
> RatingsForKPIRF <- data.frame(InputNorFor$FullName,round(InputNorFor$PredictedRankingRF,2))
> #MAE Calculation
> MAEForKpiRF <- mean(abs(TestForRF$AverageRatingOverall - TestForRF$PredictedRanking))
> MAEForKpiRF
[1] 0.6734001
```

Fig 23

Random Forest- MAE (Midfielders): The Mean Absolute Error found for the Random Forest Model of Midfielders is seen in Fig 24 below-

```
> #...................KPIs- Midfielders.............................
> #Data split- Traiing & Testing data
> TrainMidCountRF <- sample(nrow(InputNorMid), 0.8 * nrow(InputNorMid))
> TrainMidRF <- InputNorMid[, c("AverageRatingOverall","MinsPerMatch","PassesPer90","xGPer90","DuelsPer9
0","YellowCards","AerialDuelsWonPer90")][TrainMidCountRF, ]
> TestMidRF <- InputNorMid[, c("AverageRatingOverall","MinsPerMatch","PassesPer90","xGPer90","DuelsPer9
0","YellowCards","AerialDuelsWonPer90")][-TrainMidCountRF, ]
> #Training
> RF_Mid <- randomForest(AverageRatingOverall ~ ., data = TrainMidRF)
> #Rank Prediction
> TestMidRF$PredictedRanking <- predict(RF_Mid, newdata = TestMidRF)
> InputNorMid$PredictedRankingRF <- predict(RF_Mid, newdata = InputNorMid)
> RatingsMidKPIRF <- data.frame(InputNorMid$FullName,round(InputNorMid$PredictedRankingRF,2))
> #MAE Calculation
> MAEMidKpiRF <- mean(abs(TestMidRF$AverageRatingOverall - TestMidRF$PredictedRanking))
> MAEMidKpiRF
[1] 0.6347873
```

Fig 24

Random Forest- MAE (Defenders): The Mean Absolute Error found for the Random Forest Model of Defenders is seen in Fig 25 below-

```
> #...................KPIs- Defenders.............................
> #Data split- Traiing & Testing data
> TrainDefCountRF <- sample(nrow(InputNorDef), 0.8 * nrow(InputNorDef))
> TrainDefRF <- InputNorDef[, c("AverageRatingOverall","Cleansheets","YellowCards","MinsPerMatch","TacklesP
er90","CrossesPer90","ClearancesPer90","xG","AerialDuelsWon")][TrainDefCountRF, ]
> TestDefRF <- InputNorDef[, c("AverageRatingOverall","Cleansheets","YellowCards","MinsPerMatch","TacklesPe
r90","CrossesPer90","ClearancesPer90","xG","AerialDuelsWon")][-TrainDefCountRF, ]
> #Training
> RF_Def <- randomForest(AverageRatingOverall ~ ., data = TrainDefRF)
> #Rank Prediction
> TestDefRF$PredictedRanking <- predict(RF_Def, newdata = TestDefRF)
> InputNorDef$PredictedRankingRF <- predict(RF_Def, newdata = InputNorDef)
> RatingsDefKPIRF <- data.frame(InputNorDef$FullName,round(InputNorDef$PredictedRankingRF,2))
> #MAE Calculation
> MAEDefKpiRF <- mean(abs(TestDefRF$AverageRatingOverall - TestDefRF$PredictedRanking))
> MAEDefKpiRF
[1] 0.4436215
```

Fig 25

Random Forest- MAE (Goalkeepers): The Mean Absolute Error found for the Random Forest Model of Goalkeepers is seen in Fig 26 below-

```
> #...................KPIs- GKs.............................
> #Data split- Traiing & Testing data
> TrainGKCountRF <- sample(nrow(InputNorGK), 0.8 * nrow(InputNorGK))
> TrainGKRF <- InputNorGK[, c("AverageRatingOverall","Conceded","SavePercentage","MinsPerMatch","SavesPer9
0","InsideBoxSaves")][TrainGKCountRF, ]
> TestGkRF <- InputNorGK[, c("AverageRatingOverall","Conceded","SavePercentage","MinsPerMatch","SavesPer9
0","InsideBoxSaves")][-TrainGKCountRF, ]
> #Training
> RF_GK <- randomForest(AverageRatingOverall ~ ., data = TrainGKRF)
> #Rank Prediction
> TestGkRF$PredictedRanking <- predict(RF_GK, newdata = TestGkRF)
> InputNorGK$PredictedRankingRF <- predict(RF_GK, newdata = InputNorGK)
> RatingsGkKPIRF <- data.frame(InputNorGK$FullName,round(InputNorGK$PredictedRankingRF,2))
> #MAE Calculation
> MAEGkKpiRF <- mean(abs(TestGkRF$AverageRatingOverall - TestGkRF$PredictedRanking))
> MAEGkKpiRF
[1] 0.1640308
```

Fig 26

The Mean Absolute Error values for each of the models run above are collated in Table 6 below. The lower the value of the Mean Absolute Error, the better the model (Chai, T. and Draxler, R.R., 2014).

	Mean Absolute Error	
Position	**Multiple Linear Regression**	**Random Forest**
Forward	0.46 ←	0.67
Midfielder	0.44 ←	0.63
Defender	0.38 ←	0.44
Goalkeeper	0.25	0.16 ←

Table 6

4.6 Player Prediction- Machine Learning Models

The list of players collected already has an assigned Average Overall Match Rating that is provided by SofaScore based on their on-pitch performance ratings. The Average Overall Match Rating is predicted for the footballers based on their playing position by implementing machine learning models.

4.6.1 Multiple Linear Regression

The Multiple Linear Regression models were run for each of the models as described in Section 4.5.1 for each position while using the identified attributes (KPIs), and the list of top 10 predicted players in the descending order of their Average Overall Match ratings are as follows-

Forwards:

	InputNorFor.FullName	round.PredictedRatingsForKPI..2.
351	Mohamed Salah	1.34
318	Marcus Rashford	1.23
174	Harry Kane	1.18
146	Erling Haaland	1.11
14	Aleksandar Mitrovic	0.97
125	Dominic Solanke	0.93
162	Gabriel Martinelli	0.93
160	Gabriel Jesus	0.88
33	Antony	0.86
182	Heung-Min Son	0.86

Fig 27

Midfielders:

	InputNorMid.FullName	round.PredictedRatingsMidKPI..2.
387	Pascal Gross	1.16
324	Martin Odegaard	1.13
27	Andreas Pereira	1.10
210	James Maddison	1.09
60	Bruno Fernandes	1.05
63	Bukayo Saka	1.03
179	Harvey Barnes	0.98
215	James Ward-Prowse	0.96
346	Miguel Almiron	0.93
428	Ruben Neves	0.93

Fig 28

Defenders:

	InputNorDef.FullName	round.PredictedRatingsDefKPI..2.
276	Kieran Trippier	1.53
474	Trent Alexander-Arnold	1.12
147	Ethan Pinnock	1.07
268	Kenny Tete	0.93
446	Serge Aurier	0.88
369	Nayef Aguerd	0.83
468	Timothy Castagne	0.82
155	Felipe Augusto	0.79
402	Raphael Varane	0.77
48	Ben White	0.76

Fig 29

Goalkeepers:

	InputNorGK.FullName	round.PredictedRatingsGKKPI..2.
453	Stefan Ortega	2.46
38	Asmir Begovic	2.05
96	Daniel Bentley	1.53
107	David Raya	0.83
51	Bernd Leno	0.56
22	Alisson Becker	0.55
481	Vicente Guaita	0.53
106	David de Gea	0.49
375	Nick Pope	0.47
2	Aaron Ramsdale	0.40

Fig 30

4.6.2 Random Forest

The Random Forest models were run for each of the models as described in Section 4.5.2 for each position while using the identified attributes (KPIs), and the list of top 10 predicted players in the descending order of their Average Overall Match ratings are as follows-

Forwards:

	InputNorFor.FullName	round.InputNorFor.PredictedRankingRF..2.
43	Erling Haaland	0.86
45	Gabriel Martinelli	0.76
53	Ivan Toney	0.76
80	Mohamed Salah	0.73
25	Cody Gakpo	0.72
52	Heung-Min Son	0.72
7	Alexander Isak	0.70
44	Gabriel Jesus	0.70
69	Leandro Trossard	0.65
110	Willian	0.64

Fig 31

Midfielders:

	InputNorMid.FullName	round.InputNorMid.PredictedRankingRF..2.
113	Martin Odegaard	0.95
19	Bukayo Saka	0.87
156	Ruben Neves	0.87
17	Bruno Fernandes	0.86
76	James Ward-Prowse	0.86
68	Jack Grealish	0.83
153	Rodrigo Bentancur	0.82
150	Pierre-Emile Hojbjerg	0.78
21	Casemiro	0.74
38	Douglas Luiz	0.74

Fig 32

Defenders:

	InputNorDef.FullName	round.InputNorDef.PredictedRankingRF..2.
35	Diego Llorente	0.93
174	Yerry Mina	0.79
16	Ben Mee	0.74
62	James Tarkowski	0.74
42	Ethan Pinnock	0.71
152	Thiago Silva	0.71
46	Gabriel Magalhaes	0.70
44	Fabian Schar	0.68
116	Mohammed Salisu	0.68
166	Virgil van Dijk	0.65

Fig 33

Goalkeepers:

	InputNorGK.FullName	round.InputNorGK.PredictedRankingRF..2.
36	Stefan Ortega	1.52
5	Asmir Begovic	1.31
8	Daniel Bentley	0.97
12	David Raya	0.79
30	Martin Dubravka	0.65
3	Alisson Becker	0.43
6	Bernd Leno	0.37
16	Emiliano Martinez	0.33
37	Vicente Guaita	0.33
23	Jordan Pickford	0.30

Fig 34

5. Conclusion and Discussion

5.1 Discussion of the Findings

In the ever-evolving landscape of football analytics, accurately quantifying and predicting footballer performance remains a very challenging prospect. The English Premier League's 20223-2023 season data of the list of squad players of the 20 teams was taken. This research primarily delved into the intricacies of footballer performance based on his playing position.
As mentioned in Section 1.4 regarding the research questions, the study is aimed to answer the following-

i. Identify the key performance indicators (KPIs) that significantly influence a footballer's overall match rating based on his playing position on the pitch

ii. Explore the usage of machine learning models- specifically Multiple Linear Regression (MLR) and Random Forest (RF), in ranking these footballers based on the identified attributes, thereby predicting their performance ratings.

iii. Assess and compare the efficacy of these machine learning models across different player positions to determine which produces optimal predictions.

The findings of this study prove the hypotheses framed in Section 2.6.

Hypothesis H1: Different positions in football (Goalkeeper, Forward, Defender, and Midfielder) have distinct KPIs that predominantly determine their Overall Average Match Rating.

This hypothesis and the research question 1 are both proved in this study. The research reaffirmed the distinct nature that each role of football requires based on the positions- Forward, Midfielder, Defender and Goalkeeper, where each position exhibits a unique set of attributes (KPIs) that predominantly affect their overall match ratings. This was proved in Section 4.4, where the attributes were identified by passing the attributes that had passed linearity and multicollinearity tests through Multiple Linear Regression Models based on their footballer playing positions. Tables- 2, 3, 4 and 5 illustrate the key attributes for each position.

Hypothesis H2: Machine Learning algorithms- Multiple Linear Regression and Random Forest, can effectively rank footballers based on their predicted performance ratings derived from their KPIs.

Hypothesis H2 and research question 2 are both proved in this research. The study successfully implemented Multiple Linear Regression and Random Forest models to predict the average overall performance ratings of the footballers based on the identified attributes (KPIs) for each position. This was proved in Section 4.6 as both the Machine Learning Models- Multiple Linear Regression and Random Forest effectively predicted the average overall match ratings based on player positions.

Hypothesis H3: The efficacy of the Machine Learning models in predicting player performance varies based on the player's position:
H3a: For forwards, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.

H3b: For midfielders, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.
H3c: For defenders, Multiple Linear Regression will provide a more accurate prediction (lesser Mean Absolute Error) compared to Random Forest.
H3d: For goalkeepers, Random Forest will provide a more accurate prediction as it has a lesser Mean Absolute Error than that of the Multiple Linear Regression Model.

Hypothesis H3 as well as research question 3 have been proved in this dissertation. The efficacy of the models- Multiple Linear Regression and Random Forest has been conducted in Section 4.5, where the goodness of the fits of each Multiple Linear Regression model as well the Mean Absolute Errors of both the models was calculated for each footballer's position. The results of this was tabulated in Table 6 of Section 4.5, where the Hypotheses H3a, H3b, H3c and H3d have been proved.

For Forwards, Midfielders, and Defenders- Multiple Linear Regression Model exhibited superior predictive accuracy, as seen by the resulting lower Mean Absolute Error in comparison to the Random Forest Model. This suggests that there is a linear relationship between the identified attributes (KPIs) and Average Overall Match Ratings for these positions. This implies that the identified attributes (KPIs) and the Average Overall Match Ratings are directly proportional to each other, which means that an increase or decrease in a particular attribute will result in a predictable change in the Average Overall Match Rating.

For Goalkeepers- The Random Forest model employed, predicted more accurately than the Multiple Linear Regression Model, which suggests that there is a more intricate or non-linear relationship between the identified attributes (KPIs) and Average Overall Match Ratings for this position. This could be explained because the role that goalkeepers play is very unique in comparison to the other positions explored above, where they are faced with discrete events such as saves, inside the box saves or passing accuracy, which do not adhere to any linear patterns.

5.2 Implications of the Results

1. Transformation in Player Evaluation:
The findings of this study imply that there should be a shift from solely qualitative analysis of player performance to a more data-driven approach. By understanding the specific KPIs that influence player ratings, clubs can evaluate players more objectively.

2. Position-Specific Training Programs:
The identified KPIs for different positions can ensure that coaches can effectively design drills that work on improving a particular aspect of a player that affects his match rating.

3. Advanced Scouting Techniques:
Scouts can more effectively spend their time by specifically targeting players who excel in the identified KPIs that a team requires, thereby being able to recruit young gems from lesser-known clubs.

4. Business Implications:
The findings of this study could provide more impetus for even the lesser financial clubs to invest in data analytics teams to be able to compete with the rich clubs.

5. Fan Engagement:
Offering player insights and data-driven analyses derived from machine learning models could add a new dimension to fantasy football predictions, betting company odds, and player analysis segments, which could massively increase the target audience and enhance viewer engagement and help build the club's brand image.

5.3 Limitations of the Study

1.Data Constraints: This study was based on data collected from a single season of the English Premier League. The effectiveness of the accuracy of the machine learning models relies on the quality of the data. This does not take extremely unlikely events such as instances of the bottom-most team winning the league, etc. into account.

2. Model Limitations: While working with Multiple Linear Regression, there has to be a linear relationship between the dependent and independent variables and a multicollinearity test must be done as well. If these are not done, it will result in inaccurate predictions.
While working with Random Forest, since it works with non-linear relationships, it often tends to overfit the training data. This reduces the accuracy of predictions on unseen data.

Scope of KPIs: While the research identifies significant KPIs for each position, it does not take into account factors like leadership, game awareness or the mental state of a player, which can impact his performance. A footballer's performance can be influenced by external factors like refereeing decisions, weather conditions or even crowd support. These variables are difficult to quantify and integrate into models.

Positional Ambiguity: Many footballers nowadays are versatile and play multiple positions. For such players, selecting a single set of influential KPIs is challenging since a position change for a player in a one-off game could skew his performance ratings.

5.4 Future Work

1. Extended Dataset Analysis: Analyzing data from numerous seasons will lead to having a more complete knowledge of player performance trends across time. Extending the analysis to other major football leagues throughout the world may show tendencies unique to distinct playing styles or cultures.

2. Integration of Diverse Rating Systems: Evaluating footballers using ratings from many platforms would provide a more complete picture while accounting for possible biases in any single rating system.

3. Advanced Machine Learning Models: Experimenting with more advanced models such as Neural Networks or Gradient Boosted Trees may provide more accurate predictions. Cross-validation approaches could be implemented in order to lessen the risk of overfitting and increase model generalizability.

4. Incorporate Contextual Match Data: Analyzing the impact of external elements like match location (home/away), weather conditions, and even referee decisions would provide greater insight into what drives player ratings.

5. Player Versatility Analysis: With the increasing rise of versatile players who play multiple positions in modern football and with the increase in tactical flexibility, delving into these versatile players and how their attributes would differ based on different roles assigned to them will help prevent skewness of performance ratings.

6. Qualitative Insights: Combining quantitative data with qualitative analysis, such as interviews with coaches, players, or analysts, will provide a far more in-depth understanding of player performance and the possible limits of numerical player ratings.

7. Data Enrichment: Incorporating data such as player health statistics, training intensity reports, and even biometric data can assist refine our knowledge of what drives matchday performance.

6. References

- Anderson, C. and Sally, D., 2013. *The numbers game: Why everything you know about soccer is wrong*. Penguin.
- Bairner, A., 2015. Assessing the sociology of sport: On national identity and nationalism. *International Review for the Sociology of Sport*, *50*(4-5), pp.375-379.
- Barnes, C., Archer, D.T., Hogg, B., Bush, M. and Bradley, P., 2014. The evolution of physical and technical performance parameters in the English Premier League. *International journal of sports medicine*, pp.1095-1100.
- Barnes, C., Archer, D.T., Hogg, B., Bush, M. and Bradley, P., 2014. The evolution of physical and technical performance parameters in the English Premier League. *International journal of sports medicine*, pp.1095-1100.
- Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S. and Fookes, C., 2016. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), pp.2596-2605.
- Bialkowski, A., Lucey, P., Carr, P., Yue, Y. and Matthews, I., 2014, February. Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors. In *Proceedings of 8th annual MIT sloan sports analytics conference* (pp. 1-7).
- Breiman, L., 2001. Random forests. *Machine learning*, *45*, pp.5-32.
- Carling, C., Le Gall, F. and Dupont, G., 2012. Analysis of repeated high-intensity running performance in professional soccer. *Journal of sports sciences*, *30*(4), pp.325-336.
- Carmichael, F., Thomas, D. and Ward, R., 2000. Team performance: the case of English premiership football. *Managerial and decision Economics*, *21*(1), pp.31-45.
- Carmichael, F., Thomas, D. and Ward, R., 2001. Production and efficiency in association football. *Journal of sports Economics*, *2*(3), pp.228-243.
- Castañer, M., Barreira, D., Camerino, O., Anguera, M.T., Canton, A. and Hileno, R., 2016. Goal scoring in soccer: a polar coordinate analysis of motor skills used by Lionel Messi. *Frontiers in psychology*, p.806.
- Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions*, *7*(1), pp.1525-1534.
- Cheema, J.R., 2014. A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), pp.487-508.
- Cintia, P., Rinzivillo, S. and Pappalardo, L., 2015, September. A network-based approach to evaluate the performance of football teams. In *Machine learning and data mining for sports analytics workshop, Porto, Portugal*.
- Creswell, J.W. and Creswell, J.D., 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Decroos, T., Bransen, L., Van Haaren, J. and Davis, J., 2019, July. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1851-1861).
- Decroos, T., Bransen, L., Van Haaren, J. and Davis, J., 2019, July. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1851-1861).

- Dellal, A., Chamari, K., Wong, D.P., Ahmaidi, S., Keller, D., Barros, R., Bisciotti, G.N. and Carling, C., 2011. Comparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga. *European journal of sport science*, *11*(1), pp.51-59.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J. and Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), pp.27-46.
- Draper, N.R. and Smith, H., 1998. *Applied regression analysis* (Vol. 326). John Wiley & Sons.
- Duval, A. and Heerdt, D., 2020. FIFA and Human Rights-a Research Agenda. *Tilburg Law Review*, *25*(1).
- Dyte, D. and Clarke, S.R., 2000. A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research society*, *51*(8), pp.993-998.
- Eggels, H., van Elk, R. and Pechenizkiy, M., 2016. Expected goals in soccer: Explaining match results using predictive analytics. In *The machine learning and data mining for sports analytics workshop* (Vol. 16).
- Fernández, J., Bornn, L. and Cervone, D., 2021. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, *110*(6), pp.1389-1427.
- Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R. and McRobert, A.P., 2016. Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of sports sciences*, *34*(24), pp.2195-2204.
- Field, A., 2013. *Discovering statistics using IBM SPSS statistics*. sage.
- FIFA, 2014a. *2014 FIFA world cup reached 3.2 billion viewers; 1 billion watched final*. [Online][2015]. Sports Business Journal. Available from: https://www.sportsbusinessjournal.com/Global/Issues/2015/12/17/Media/FIFA-World-Cup.aspx
- Goldblatt, D., 2007. *The ball is round: a global history of football*. Penguin UK.
- Grund, T.U., 2012. Network structure and team performance: The case of English Premier League soccer teams. *Social Networks*, *34*(4), pp.682-690.
- Gudmundsson, J. and Horton, M., 2017. Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, *50*(2), pp.1-34.
- Gyarmati, L. and Stanojevic, R., 2016. Qpass: a merit-based evaluation of soccer passes. *arXiv preprint arXiv:1608.03532*.
- He, M., Cachucho, R. and Knobbe, A.J., 2015, June. Football Player's Performance and Market Value. In *Mlsa@ pkdd/ecml* (pp. 87-95).
- Hughes, M. and Franks, I., 2005. Analysis of passing sequences, shots and goals in soccer. *Journal of sports sciences*, *23*(5), pp.509-514.
- Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), pp.679-688.
- James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
- James, N., 2006. Notational analysis in soccer: past, present and future. *International journal of performance analysis in sport*, *6*(2), pp.67-81.

- Ji, S., Li, Q., Cao, W., Zhang, P. and Muccini, H., 2020. Quality assurance technologies of big data applications: A systematic literature review. *Applied Sciences*, *10*(22), p.8052.
- Jolliffe, I.T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), p.20150202.
- Jones, P.D., James, N. and Mellalieu, S.D., 2004. Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport*, *4*(1), pp.98-102.
- Karlis, D. and Ntzoufras, I., 2003. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), pp.381-393.
- King, A., 2017. *The European ritual: Football in the new Europe*. Routledge.
- Kreft, L., 2014. Aesthetics of the beautiful game. *Soccer & society*, *15*(3), pp.353-375.
- Lago-Peñas, C. and Dellal, A., 2010. Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables. *Journal of human kinetics*, *25*(2010), pp.93-100.
- Lewis, M., 2004. *Moneyball: The art of winning an unfair game*. WW Norton & Company.
- Liu, H., Gómez, M.A., Gonçalves, B. and Sampaio, J., 2016. Technical performance and match-to-match variation in elite football teams. *Journal of sports sciences*, *34*(6), pp.509-518.
- Liu, H., Gómez, M.A., Gonçalves, B. and Sampaio, J., 2016. Technical performance and match-to-match variation in elite football teams. *Journal of sports sciences*, *34*(6), pp.509-518.
- Liu, H., Gomez, M.Á., Lago-Peñas, C. and Sampaio, J., 2015. Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *Journal of sports sciences*, *33*(12), pp.1205-1213.
- Lucey, P., Oliver, D., Carr, P., Roth, J. and Matthews, I., 2013, August. Assessing team strategy using spatiotemporal data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1366-1374).
- Merriam, S.B. and Tisdell, E.J., 2015. *Qualitative research: A guide to design and implementation*. John Wiley & Sons.
- Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. *Introduction to linear regression analysis*. John Wiley & Sons.
- Moroney, M.J., 1956. *Facts from figures* (Vol. 236). Harmondsworth, Middlesex: Penguin books.
- O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, *41*, pp.673-690.
- Osborne, J.W. and Overbay, A., 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), p.6.
- Palacios-Huerta, I., 2004. Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications*, *13*, pp.241-258.
- Palacios-Huerta, I., 2014. *Beautiful game theory: How soccer can help economics*. Princeton University Press.

- Patro, S.G.O.P.A.L. and Sahu, K.K., 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.
- Peng, K., Cooke, J., Crockett, A., Shin, D., Foster, A., Rue, J., Williams, R., Valeiras, J., Scherer, W., Tuttle, C. and Adams, S., 2018, April. Predictive analytics for University of Virginia football recruiting. In *2018 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 243-248). IEEE.
- Perarnau, M., 2014. *Pep Confidential: Inside Pep Guardiola's First Season at Bayern Munich*. Birlinn.
- Pifer, N.D., Wang, Y., Scremin, G., Pitts, B.G. and Zhang, J.J., 2018. Contemporary global football industry: an introduction. *The Global Football Industry*, pp.3-35.
- Plumley, D.J., Wilson, R. and Shibli, S., 2017. A holistic performance assessment of English Premier League football clubs 1992-2013. *Journal of Applied Sport Management*, *9*(1).
- Pollard, R. and Reep, C., 1997. Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society Series D: The Statistician*, *46*(4), pp.541-550.
- Power, P., Ruiz, H., Wei, X. and Lucey, P., 2017, August. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1605-1613).
- Reep, C. and Benjamin, B., 1968. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, *131*(4), pp.581-585.
- Rein, R. and Memmert, D., 2016. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), pp.1-13.
- Rein, R. and Memmert, D., 2016. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), pp.1-13.
- Ridder, G., Cramer, J.S. and Hopstaken, P., 1994. Down to ten: Estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, *89*(427), pp.1124-1127.
- Ridder, G., Cramer, J.S. and Hopstaken, P., 1994. Down to ten: Estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, *89*(427), pp.1124-1127.
- Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernández, J. and Medina, D., 2018. Effective injury forecasting in soccer with GPS training data and machine learning. *PloS one*, *13*(7), p.e0201264.
- Ruijg, J. and van Ophem, H., 2015. Determinants of football transfers. *Applied Economics Letters*, *22*(1), pp.12-19.
- Sampaio, J., Lago, C. and Drinkwater, E.J., 2010. Explanations for the United States of America's dominance in basketball at the Beijing Olympic Games (2008). *Journal of Sports Sciences*, *28*(2), pp.147-152.
- Sarmento, H., Pereira, A., Matos, N., Campaniço, J., Anguera, T.M. and Leitão, J., 2013. English premier league, spaiṅs la liga and italýs seriés a–What's different?. *International Journal of Performance Analysis in Sport*, *13*(3), pp.773-789.
- Schober, P., Boer, C. and Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), pp.1763-1768.

- Scrucca, L. and Serafini, A., 2019. Projection pursuit based on Gaussian mixtures and evolutionary algorithms. *Journal of Computational and Graphical Statistics*, *28*(4), pp.847-860.
- Shaw, L. and Glickman, M., 2019. Dynamic analysis of team strategy in professional football. *Barça sports analytics summit*, *13*.
- Statista Research Department. 2023. *Revenue of the leading European soccer leagues 2012-2024, by league*. [Online] [Accessed September 11, 2023] Available from: https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue/
- Taylor, J.B., Mellalieu, S.D. and James, N., 2004. Behavioural comparisons of positional demands in professional soccer. *International Journal of Performance Analysis in Sport*, *4*(1), pp.81-97.
- Tobar, F. and Ramshaw, G., 2022. 'Welcome to the EPL': analysing the development of football tourism in the English Premier League. *Soccer & Society*, *23*(4-5), pp.432-450.
- Tomlinson, A. and Young, C. eds., 2006. *National identity and global sports events: Culture, politics, and spectacle in the Olympics and the football World Cup*. Suny Press.
- UEFA, 2023. *Country coefficients*. Uefa.com. [Online] [2023] Available from: https://www.uefa.com/nationalassociations/uefarankings/country/#/yr/2024
- Vöpel, H., 2011. Do we really need financial fair play in european club football? an economic analysis. *CESifo DICE Report*, *9*(3), pp.54-59.
- Vrooman, J., 2007. Theory of the beautiful game: The unification of European football. *Scottish journal of political economy*, *54*(3), pp.314-354.
- Williams, M.N., Grajales, C.A.G. and Kurkiewicz, D., 2013. Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, *18*(1), p.11.
- Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), pp.79-82.
- Yiannakos, A. and Armatas, V., 2006. Evaluation of the goal scoring patterns in European Championship in Portugal 2004. *International Journal of Performance Analysis in Sport*, *6*(1), pp.178-188.
- Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q.P. and Lillard Jr, J.W., 2014. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, *4*(5), p.9.

7. Appendix A- R Code

```
#Installing necessary packages
install.packages("gmodels")
install.packages("NbClust")
install.packages('caret')
install.packages("tidyr")
install.packages("fastcluster")
install.packages("ggcorrplot")
install.packages('e1071')
install.packages("mice")
install.packages("dendextend")
install.packages("Hmisc")
install.packages('C50')
install.packages('tidyverse')
install.packages("gridExtra")
install.packages('e1071')
install.packages("ggplot2")
install.packages('pROC')
install.packages("VIM")
install.packages("randomForest")
install.packages("class")
install.packages("stringr")
install.packages("dplyr")
install.packages('ROCR')
install.packages("corrplot")
install.packages("ggrepel")

library(ROCR)
library(tidyr)
library(ggcorrplot)
library(ggrepel)
library(C50)
library(stringr)
library(gridExtra)
library(NbClust)
library(e1071)
library(dplyr)
library(caret)
library(tidyverse)
library(e1071)
library(corrplot)
library(ggplot2)
library(dendextend)
library(pROC)
library(fastcluster)
library(VIM)
library(class)
library(gmodels)
library(randomForest)
```

```
library(mice)
library(Hmisc)

rm(list=ls())

#Loading the csv
InputData <- read.csv("EPL22-23.csv",header = TRUE,stringsAsFactors = T, encoding =
'UTF-8')

#Checking the datatype of each Variable
str(InputData)

#Overall distribution of EPL footballers stats
summary(InputData)
table(InputData$Position)

#Missing values check
md.pattern(InputData, rotate.names = TRUE)

#Outliers Check
outliers <- function(InputData)
{
  InputData %>%  select_if(is.numeric) %>% map(~ boxplot.stats(.x)$out)
}
outliers(InputData)

#-------------------------DESCRIPTIVE STATISTICS----------------------

#-------------------------FORWARDS-----------------------------------

# Filter forwards who played more than 500 minutes and rank by criteria
forwards_ranked <- InputData %>%
  filter(MinutesPlayed > 500 & Position == "Forward") %>%
  arrange(desc(Shots), desc(DuelsPer90))

# Select top 15 forwards
top_15_forwards <- head(forwards_ranked, 15)

# Create a scatter plot for DuelsPer90 vs. Shots for the top 15 forwards
ggplot(top_15_forwards, aes(x = DuelsPer90, y = Shots, label = FullName)) +
  geom_point(color = "blue", size = 4) +
  geom_text_repel(size = 3.5, nudge_y = 1.5) +
  theme_minimal() +
  labs(title = "DuelsPer90 vs. Shots for Top 15 Forwards",
       x = "Duels Per 90",
       y = "Number of Shots")


# Filter the data for forwards who played more than 500 minutes
forwards_ranked <- InputData %>%
```

```
  filter(MinutesPlayed > 500 & Position == "Forward") %>%
  arrange(desc(Dribbles), desc(KeyPasses))  # Sort by Dribbles and then by KeyPasses

# Select the top 15 forwards
top_15_forwards <- head(forwards_ranked, 15)

# Create the scatter plot for Dribbles vs. KeyPasses
ggplot(top_15_forwards, aes(x = Dribbles, y = KeyPasses)) +
  geom_point(color = "blue", size = 5, alpha = 0.7) +  # Adjust color, size, and transparency
of points if needed
  geom_text_repel(aes(label = FullName), nudge_x = 0.5) +  # Use ggrepel to prevent text
overlapping
  theme_minimal() +
  labs(title = "Dribbles vs. KeyPasses for Top 15 Forwards",
       x = "Dribbles",
       y = "KeyPasses")

#-------------------------MIDFIELDERS----------------------------------
# Filter midfielders who played more than 500 minutes and rank by criteria
midfielders_ranked <- InputData %>%
  filter(MinutesPlayed > 500 & Position == "Midfielder") %>%
  arrange(desc(xA), desc(Crosses))

# Select top 15 midfielders
top_15_midfielders <- head(midfielders_ranked, 15)

# Create a scatter plot for Crosses vs. xA for the top 15 midfielders
ggplot(top_15_midfielders, aes(x = Crosses, y = xA, label = FullName)) +
  geom_point(color = "red", size = 4) +
  geom_text_repel(size = 3.5, nudge_y = 0.5) +
  theme_minimal() +
  labs(title = "Crosses vs. xA for Top 15 Midfielders",
       x = "Number of Crosses",
       y = "xA (Expected Assists)")

#-------------------------DEFENDERS----------------------------------
# Filter defenders who played more than 500 minutes
defenders_ranked <- InputData %>%
  filter(MinutesPlayed > 500 & Position == "Defender") %>%
  arrange(desc(Tackles), desc(Passes))

# Select top 10 defenders
top_10_defenders <- head(defenders_ranked, 10)

# Create a bubble chart for Passes vs. Tackles
ggplot(top_10_defenders, aes(x = Passes, y = Tackles)) +
  geom_point(aes(color = FullName), size = 5, alpha = 0.6) +  # Setting alpha for a bit of
transparency
  scale_color_brewer(palette = "Set3") +  # A different color palette for distinction
```

```
  geom_text_repel(aes(label = FullName), size = 4, nudge_x = 0.5) +  # You might need to
adjust nudge_x to ensure text doesn't overlap too much
  theme_minimal() +
  labs(title = "Tackles vs. Passes for Top 10 Defenders",
       x = "Passes",
       y = "Number of Tackles")


#-------------------------GOALKEEPERS----------------------------------
# Filter goalkeepers who played more than 500 minutes
goalkeepers_ranked <- InputData %>%
  filter(MinutesPlayed > 500 & Position == "Goalkeeper") %>%
  arrange(desc(SavePercentage), desc(PassCompletionRate))

# Select top 15 goalkeepers
top_15_goalkeepers <- head(goalkeepers_ranked, 15)

# Create a bubble chart for PassCompletionRate vs. SavePercentage
ggplot(top_15_goalkeepers, aes(x = PassCompletionRate, y = SavePercentage)) +
  geom_point(color = "orange", size = 4, alpha = 0.6) +  # All points will be orange and of
the same size
  geom_text_repel(aes(label = FullName), nudge_x = 0.5) +  # Use ggrepel to prevent text
overlapping
  theme_minimal() +
  labs(title = "Save Percentage vs. Pass Completion Rate for Top 15 Goalkeepers",
       x = "Pass Completion Rate",
       y = "Save Percentage")

#-----------------------------------------------------------------------
#Datatype conversion of columns to numeric
Input_DataType <- InputData %>% mutate_at(c('MinutesPlayed',
                                    'Appearances',
                                    'Goals',
                                    'Assists',
                                    'PenaltyGoals',
                                    'PenaltyMisses',
                                    'Cleansheets',
                                    'Conceded',
                                    'YellowCards',
                                    'RedCards',
                                    'MinsPerMatch',
                                    'AverageRatingOverall',
                                    'PassesPer90',
                                    'Passes',
                                    'PassesCompletedPer90',
                                    'KeyPassesPer90',
                                    'KeyPasses',
                                    'TacklesPer90',
                                    'Tackles',
                                    'ShotsPer90',
```

```
                                'Shots',
                                'ShotsOnTarget',
                                'ShotsOnTargetPer90',
                                'Interceptions',
                                'InterceptionsPer90',
                                'Crosses',
                                'CrossesPer90',
                                'Dribbles',
                                'DribblesPer90',
                                'DribblesSuccessful',
                                'DribblesSuccessfulPer90',
                                'Saves',
                                'SavesPer90',
                                'ShotsFaced',
                                'ShotsFacedPer90',
                                'SavePercentage',
                                'xG',
                                'PassCompletionRate',
                                'DribbledPastPer90',
                                'DribbledPast',
                                'InsideBoxSaves',
                                'BlocksPer90',
                                'Blocks',
                                'ClearancesPer90',
                                'Clearances',
                                'PenaltiesCommitted',
                                'Punches',
                                'PunchesPer90',
                                'Offisdes',
                                'xA',
                                'xAPer90',
                                'nPxG',
                                'xPxGPer90',
                                'FoulsDrawn',
                                'FoulsDrawnPer90',
                                'FoulsCommittedPer90',
                                'FoulsCommitted',
                                'xGPer90',
                                'AerialDuelsWon',
                                'AerialDuelsWonPer90',
                                'DuelsPer90',
                                'Duels',
                                'DuelsWon',
                                'DuelsWonPer90',
                                'Dispossesed',
                                'DispossesedPer90',
                                'AccurateCrosses',
                                'AccurateCrossesPer90'), as.numeric)

str(Input_DataType)
```

```
#Variables for which lesser value is considered to be better, are inversed
Inv = c('PenaltyMisses',
        'Conceded',
        'YellowCards',
        'RedCards',
        'DribbledPast',
        'PenaltiesCommitted',
        'Offisdes',
        'FoulsCommittedPer90',
        'FoulsCommitted',
        'Dispossesed',
        'DispossesedPer90')

Input_DataType[Inv] = 1/Input_DataType[Inv]
Input_DataType <- replace(Input_DataType, Input_DataType==Inf, 0)

#Creating subsets based on positions
Input_DataFor <- subset(Input_DataType, Position == "Forward")
Input_DataMid <- subset(Input_DataType, Position == "Midfielder")
Input_DataDef <- subset(Input_DataType, Position == "Defender")
Input_DataGoal <- subset(Input_DataType, Position == "Goalkeeper")

#Normalizing values of variables to a single scale
Nor_ColsFor <- c('MinutesPlayed',
                 'Appearances',
                 'Goals',
                 'Assists',
                 'PenaltyMisses',
                 'YellowCards',
                 'RedCards',
                 'MinsPerMatch',
                 'AverageRatingOverall',
                 'PassesPer90',
                 'Passes',
                 'PassesCompletedPer90',
                 'KeyPassesPer90',
                 'KeyPasses',
                 'ShotsPer90',
                 'Shots',
                 'ShotsOnTarget',
                 'ShotsOnTargetPer90',
                 'Interceptions',
                 'InterceptionsPer90',
                 'Crosses',
                 'CrossesPer90',
                 'Dribbles',
                 'DribblesPer90',
                 'DribblesSuccessful',
                 'DribblesSuccessfulPer90',
```

```
                'xG',
                'Offisdes',
                'xA',
                'xAPer90',
                'nPxG',
                'xPxGPer90',
                'FoulsDrawn',
                'FoulsDrawnPer90',
                'xGPer90',
                'AerialDuelsWon',
                'AerialDuelsWonPer90',
                'DuelsPer90',
                'Duels',
                'DuelsWon',
                'DuelsWonPer90')

Nor_ColsMid <- c('MinutesPlayed',
                'Appearances',
                'Goals',
                'Assists',
                'PenaltyMisses',
                'YellowCards',
                'RedCards',
                'MinsPerMatch',
                'AverageRatingOverall',
                'PassesPer90',
                'Passes',
                'PassesCompletedPer90',
                'KeyPassesPer90',
                'KeyPasses',
                'TacklesPer90',
                'Tackles',
                'ShotsPer90',
                'Interceptions',
                'InterceptionsPer90',
                'Crosses',
                'CrossesPer90',
                'Dribbles',
                'DribblesPer90',
                'DribblesSuccessful',
                'DribblesSuccessfulPer90',
                'xG',
                'PassCompletionRate',
                'BlocksPer90',
                'Blocks',
                'PenaltiesCommitted',
                'xA',
                'xAPer90',
                'nPxG',
                'xPxGPer90',
```

```
                'FoulsDrawn',
                'FoulsDrawnPer90',
                'xGPer90',
                'AerialDuelsWon',
                'AerialDuelsWonPer90',
                'DuelsPer90',
                'Duels',
                'DuelsWon',
                'DuelsWonPer90',
                'Dispossesed',
                'DispossesedPer90',
                'AccurateCrosses',
                'AccurateCrossesPer90')

Nor_ColsDef <- c('MinutesPlayed',
                'Appearances',
                'Goals',
                'Assists',
                'Cleansheets',
                'Conceded',
                'YellowCards',
                'RedCards',
                'MinsPerMatch',
                'AverageRatingOverall',
                'PassesPer90',
                'Passes',
                'PassesCompletedPer90',
                'KeyPassesPer90',
                'KeyPasses',
                'TacklesPer90',
                'Tackles',
                'ShotsPer90',
                'Shots',
                'Interceptions',
                'InterceptionsPer90',
                'Crosses',
                'CrossesPer90',
                'Dribbles',
                'DribblesPer90',
                'DribblesSuccessful',
                'DribblesSuccessfulPer90',
                'xG',
                'PassCompletionRate',
                'DribbledPastPer90',
                'DribbledPast',
                'BlocksPer90',
                'Blocks',
                'ClearancesPer90',
                'Clearances',
                'PenaltiesCommitted',
```

```
                'xA',
                'xAPer90',
                'nPxG',
                'xPxGPer90',
                'FoulsCommittedPer90',
                'FoulsCommitted',
                'xGPer90',
                'AerialDuelsWon',
                'AerialDuelsWonPer90',
                'DuelsPer90',
                'Duels',
                'DuelsWon',
                'DuelsWonPer90',
                'Dispossesed',
                'DispossesedPer90',
                'AccurateCrosses',
                'AccurateCrossesPer90')

Nor_ColsGoal <- c('MinutesPlayed',
                'Appearances',
                'Cleansheets',
                'Conceded',
                'YellowCards',
                'RedCards',
                'MinsPerMatch',
                'AverageRatingOverall',
                'Passes',
                'KeyPasses',
                'Interceptions',
                'Saves',
                'SavesPer90',
                'ShotsFaced',
                'ShotsFacedPer90',
                'SavePercentage',
                'PassCompletionRate',
                'InsideBoxSaves',
                'ClearancesPer90',
                'Clearances',
                'PenaltiesCommitted',
                'Punches',
                'PunchesPer90',
                'FoulsCommittedPer90',
                'FoulsCommitted',
                'AerialDuelsWon',
                'AerialDuelsWonPer90',
                'DuelsPer90',
                'Duels',
                'DuelsWon',
                'DuelsWonPer90')
```

```
NorDataFor <- as.data.frame(scale(Input_DataFor[Nor_ColsFor]))
NorDataMid <- as.data.frame(scale(Input_DataMid[Nor_ColsMid]))
NorDataDef <- as.data.frame(scale(Input_DataDef[Nor_ColsDef]))
NorDataGK <- as.data.frame(scale(Input_DataGoal[Nor_ColsGoal]))

#Normalized values combined with EPL Footballer Data
InputNorFor <- cbind(Input_DataFor$FullName,Input_DataFor$Position,NorDataFor)
InputNorMid <- cbind(Input_DataMid$FullName,Input_DataMid$Position,NorDataMid)
InputNorDef <- cbind(Input_DataDef$FullName,Input_DataDef$Position,NorDataDef)
InputNorGK <- cbind(Input_DataGoal$FullName,Input_DataGoal$Position,NorDataGK)

colnames(InputNorFor)[1] ="FullName"
colnames(InputNorMid)[1] ="FullName"
colnames(InputNorDef)[1] ="FullName"
colnames(InputNorGK)[1] ="FullName"

#Correlation check- Forwards
ResFor <- InputNorFor[,c('Appearances',
                         'PenaltyMisses',
                         'YellowCards',
                         'RedCards',
                         'MinsPerMatch',
                         'KeyPassesPer90',
                         'ShotsPer90',
                         'ShotsOnTargetPer90',
                         'InterceptionsPer90',
                         'CrossesPer90',
                         'DribblesSuccessfulPer90',
                         'xG',
                         'Offisdes',
                         'xA',
                         'xGPer90',
                         'AerialDuelsWonPer90',
                         'DuelsPer90',
                         'DuelsWonPer90')]

round(ResFor,2)

#Calculating the Correlation
CorrFor= round(cor(ResFor), 2)

#Display Findings
ggcorrplot(CorrFor, hc.order = TRUE, type = "lower",
           lab = TRUE)

#Correlation check- Midfielders
ResMid <- InputNorMid[,c('Appearances',
                         'Assists',
                         'YellowCards',
                         'RedCards',
```

```
                    'MinsPerMatch',
                    'PassesPer90',
                    'KeyPassesPer90',
                    'TacklesPer90',
                    'ShotsPer90',
                    'InterceptionsPer90',
                    'CrossesPer90',
                    'DribblesPer90',
                    'DribblesSuccessful',
                    'xG',
                    'PassCompletionRate',
                    'BlocksPer90',
                    'xAPer90',
                    'FoulsDrawnPer90',
                    'xGPer90',
                    'AerialDuelsWonPer90',
                    'DuelsPer90',
                    'DuelsWonPer90')]

round(ResMid,2)

#Calculating the Correlation
CorrMid = round(cor(ResMid), 2)

#Display Findings
ggcorrplot(CorrMid, hc.order = TRUE, type = "lower",
        lab = TRUE)

#Correlation check- Defenders
ResDef <- InputNorDef[,c('Appearances',
                    'Assists',
                    'Cleansheets',
                    'Conceded',
                    'YellowCards',
                    'RedCards',
                    'MinsPerMatch',
                    'PassesPer90',
                    'KeyPassesPer90',
                    'TacklesPer90',
                    'ShotsPer90',
                    'InterceptionsPer90',
                    'CrossesPer90',
                    'DribblesPer90',
                    'PassCompletionRate',
                    'DribbledPastPer90',
                    'BlocksPer90',
                    'ClearancesPer90',
                    'xAPer90',
                    'FoulsCommittedPer90',
                    'xG',
```

```
                    'AerialDuelsWon',
                    'Duels',
                    'DuelsWonPer90',
                    'DispossesedPer90')]

round(ResDef,2)

#Calculating the Correlation
CorrDef = round(cor(ResDef), 2)

#Display Findings
ggcorrplot(CorrDef, hc.order = TRUE, type = "lower",
        lab = TRUE)

#Correlation check- Goalkeepers
ResGK <- InputNorGK[,c('Cleansheets',
                    'Conceded',
                    'YellowCards',
                    'RedCards',
                    'MinsPerMatch',
                    'KeyPasses',
                    'Interceptions',
                    'SavesPer90',
                    'ShotsFacedPer90',
                    'SavePercentage',
                    'PassCompletionRate',
                    'InsideBoxSaves',
                    'ClearancesPer90',
                    'FoulsCommitted',
                    'AerialDuelsWonPer90',
                    'DuelsPer90')]

round(ResGK,2)

#Calculating the Correlation
CorrGK = round(cor(ResGK), 2)

#Display Findings
ggcorrplot(CorrGK, hc.order = TRUE, type = "lower",
        lab = TRUE)

#KPI Identification: Position-wise
#Multiple Linear Regression- Forwards
MLRFor = lm(formula =  AverageRatingOverall ~ Appearances +
            PenaltyMisses +
            YellowCards +
            RedCards +
            MinsPerMatch +
            KeyPassesPer90 +
            ShotsPer90 +
```

```
            ShotsOnTargetPer90 +
            InterceptionsPer90 +
            CrossesPer90 +
            DribblesSuccessfulPer90 +
            xG +
            Offisdes +
            xA +
            xGPer90 +
            AerialDuelsWonPer90 +
            DuelsPer90 +
            DuelsWonPer90, data = InputNorFor)

#Display results
summary(MLRFor)

PredictedRatingsFor <- predict(MLRFor, newdata = ResFor)
PredictedRatingsFor

RatingsFor <- data.frame(InputNorFor$FullName,round(PredictedRatingsFor,2))

#Multiple Linear Regression- Midfielders
MLRMid = lm(formula =  AverageRatingOverall ~ Appearances +
            Assists +
            YellowCards +
            RedCards +
            MinsPerMatch +
            PassesPer90 +
            KeyPassesPer90 +
            TacklesPer90 +
            ShotsPer90 +
            InterceptionsPer90 +
            CrossesPer90 +
            DribblesPer90 +
            DribblesSuccessful +
            xG +
            PassCompletionRate +
            BlocksPer90 +
            xAPer90 +
            FoulsDrawnPer90 +
            xGPer90 +
            AerialDuelsWonPer90 +
            DuelsPer90, data = InputNorMid)

#Display results
summary(MLRMid)

PredictedRatingsMid <- predict(MLRMid, newdata = ResMid)
PredictedRatingsMid

RatingsMid <- data.frame(InputNorMid$FullName,round(PredictedRatingsMid,2))
```

```
#Multiple Linear Regression- Defenders
MLRDef = lm(formula =  AverageRatingOverall ~ Assists +
            Cleansheets +
            Conceded +
            YellowCards +
            RedCards +
            MinsPerMatch +
            PassesPer90 +
            KeyPassesPer90 +
            TacklesPer90 +
            ShotsPer90 +
            InterceptionsPer90 +
            CrossesPer90 +
            DribblesPer90 +
            PassCompletionRate +
            DribbledPastPer90 +
            BlocksPer90 +
            ClearancesPer90 +
            xAPer90 +
            FoulsCommittedPer90 +
            xG +
            AerialDuelsWon +
            DuelsWonPer90 +
            DispossesedPer90, data = InputNorDef)

#Display results
summary(MLRDef)

PredictedRatingsDef <- predict(MLRDef, newdata = ResDef)
PredictedRatingsDef

RatingsDef <- data.frame(InputNorDef$FullName,round(PredictedRatingsDef,2))

#Multiple Linear Regression- GK
MLRGK = lm(formula =  AverageRatingOverall ~ Cleansheets +
             Conceded +
             YellowCards +
             RedCards +
             MinsPerMatch +
             KeyPasses +
             Interceptions +
             SavesPer90 +
             SavePercentage +
             PassCompletionRate +
             InsideBoxSaves +
             ClearancesPer90 +
             FoulsCommitted +
             AerialDuelsWonPer90 +
             DuelsPer90, data = InputNorGK)
```

```
#Display results
summary(MLRGK)

PredictedRatingsGK <- predict(MLRGK, newdata = ResGK)
PredictedRatingsGK

RatingsGK <- data.frame(InputNorGK$FullName,round(PredictedRatingsGK,2))

#.....................Multiple Linear Regression.....................

#Significant KPIs- Forwards
MLRForKPI = lm(formula =  AverageRatingOverall ~ MinsPerMatch +
               ShotsPer90 +
               DribblesSuccessfulPer90 +
               DuelsPer90 +
               DuelsWonPer90, data = InputNorFor)

#Display results
summary(MLRForKPI)

PredictedRatingsForKPI <- predict(MLRForKPI, newdata = ResFor)
PredictedRatingsForKPI

RatingsForKPI <- data.frame(InputNorFor$FullName,round(PredictedRatingsForKPI,2))

MAEForKPI <- mean(abs(InputNorFor$AverageRatingOverall -
PredictedRatingsForKPI))
MAEForKPI

#Significant KPIs- Midfielders
MLRMidKPI = lm(formula =  AverageRatingOverall ~ MinsPerMatch +
               PassesPer90 +
               xGPer90 +
               DuelsPer90 + YellowCards +
               AerialDuelsWonPer90, data = InputNorMid)

#Display results
summary(MLRMidKPI)

PredictedRatingsMidKPI <- predict(MLRMidKPI, newdata = ResMid)
PredictedRatingsMidKPI

RatingsMidKPI <-
data.frame(InputNorMid$FullName,round(PredictedRatingsMidKPI,2))

MAEMidKPI <- mean(abs(InputNorMid$AverageRatingOverall -
PredictedRatingsMidKPI))
MAEMidKPI
```

```
#Significant KPIs- Defenders
MLRDefKPI = lm(formula =  AverageRatingOverall ~ Cleansheets +
               YellowCards +
               MinsPerMatch +
               TacklesPer90 +
               CrossesPer90 +
               ClearancesPer90 +
               xG +
               AerialDuelsWon, data = InputNorDef)

#Display results
summary(MLRDefKPI)

PredictedRatingsDefKPI <- predict(MLRDefKPI, newdata = ResDef)
PredictedRatingsDefKPI

RatingsDefKPI <- data.frame(InputNorDef$FullName,round(PredictedRatingsDefKPI,2))

MAEDefKPI <- mean(abs(InputNorDef$AverageRatingOverall -
PredictedRatingsDefKPI))
MAEDefKPI

#Significant KPIs- GKs
MLRGkKPI = lm(formula =  AverageRatingOverall ~ Conceded +
               SavePercentage + SavesPer90 + InsideBoxSaves +
               MinsPerMatch, data = InputNorGK)

#Display results
summary(MLRGkKPI)

PredictedRatingsGKKPI <- predict(MLRGkKPI, newdata = ResGK)
PredictedRatingsGKKPI

RatingsGkKPI <- data.frame(InputNorGK$FullName,round(PredictedRatingsGKKPI,2))

MAEGkKPI <- mean(abs(InputNorGK$AverageRatingOverall -
PredictedRatingsGKKPI))
MAEGkKPI

#.....Random Forest.....

#.....KPIs- Forwards.....
#Data split- Traiing & Testing data

set.seed(12345)
TrainForCountRF <- sample(nrow(InputNorFor), 0.8 * nrow(InputNorFor))
TrainForRF <- InputNorFor[,
c("AverageRatingOverall","MinsPerMatch","ShotsPer90","DribblesSuccessfulPer90","D
uelsPer90","DuelsWonPer90")][TrainForCountRF, ]
```

```
TestForRF <- InputNorFor[, c("AverageRatingOverall","MinsPerMatch","ShotsPer90","DribblesSuccessfulPer90","DuelsPer90","DuelsWonPer90")][-TrainForCountRF, ]

#Training
RF_For <- randomForest(AverageRatingOverall ~ ., data = TrainForRF)

#Rank Prediction
TestForRF$PredictedRanking <- predict(RF_For, newdata = TestForRF)
InputNorFor$PredictedRankingRF <- predict(RF_For, newdata = InputNorFor)

RatingsForKPIRF <- data.frame(InputNorFor$FullName,round(InputNorFor$PredictedRankingRF,2))

#MAE Calculation
MAEForKpiRF <- mean(abs(TestForRF$AverageRatingOverall - TestForRF$PredictedRanking))
MAEForKpiRF

#.....KPIs- Midfielders.....
#Data split- Traiing & Testing data
TrainMidCountRF <- sample(nrow(InputNorMid), 0.8 * nrow(InputNorMid))
TrainMidRF <- InputNorMid[, c("AverageRatingOverall","MinsPerMatch","PassesPer90","xGPer90","DuelsPer90","YellowCards","AerialDuelsWonPer90")][TrainMidCountRF, ]
TestMidRF <- InputNorMid[, c("AverageRatingOverall","MinsPerMatch","PassesPer90","xGPer90","DuelsPer90","YellowCards","AerialDuelsWonPer90")][-TrainMidCountRF, ]

#Training
RF_Mid <- randomForest(AverageRatingOverall ~ ., data = TrainMidRF)

#Rank Prediction
TestMidRF$PredictedRanking <- predict(RF_Mid, newdata = TestMidRF)
InputNorMid$PredictedRankingRF <- predict(RF_Mid, newdata = InputNorMid)

RatingsMidKPIRF <- data.frame(InputNorMid$FullName,round(InputNorMid$PredictedRankingRF,2))

#MAE Calculation
MAEMidKpiRF <- mean(abs(TestMidRF$AverageRatingOverall - TestMidRF$PredictedRanking))
MAEMidKpiRF

#.....KPIs- Defenders.....
#Data split- Traiing & Testing data
TrainDefCountRF <- sample(nrow(InputNorDef), 0.8 * nrow(InputNorDef))
TrainDefRF <- InputNorDef[, c("AverageRatingOverall","Cleansheets","YellowCards","MinsPerMatch","TacklesPer90","CrossesPer90","ClearancesPer90","xG","AerialDuelsWon")][TrainDefCountRF, ]
```

```
TestDefRF <- InputNorDef[, c("AverageRatingOverall","Cleansheets","YellowCards","MinsPerMatch","TacklesPer90","CrossesPer90","ClearancesPer90","xG","AerialDuelsWon")][-TrainDefCountRF, ]

#Training
RF_Def <- randomForest(AverageRatingOverall ~ ., data = TrainDefRF)

#Rank Prediction
TestDefRF$PredictedRanking <- predict(RF_Def, newdata = TestDefRF)
InputNorDef$PredictedRankingRF <- predict(RF_Def, newdata = InputNorDef)

RatingsDefKPIRF <- data.frame(InputNorDef$FullName,round(InputNorDef$PredictedRankingRF,2))

#MAE Calculation
MAEDefKpiRF <- mean(abs(TestDefRF$AverageRatingOverall - TestDefRF$PredictedRanking))
MAEDefKpiRF

#.....KPIs- GKs.....
#Data split- Traiing & Testing data
TrainGKCountRF <- sample(nrow(InputNorGK), 0.8 * nrow(InputNorGK))
TrainGKRF <- InputNorGK[, c("AverageRatingOverall","Conceded","SavePercentage","MinsPerMatch","SavesPer90","InsideBoxSaves")][TrainGKCountRF, ]
TestGkRF <- InputNorGK[, c("AverageRatingOverall","Conceded","SavePercentage","MinsPerMatch","SavesPer90","InsideBoxSaves")][-TrainGKCountRF, ]

#Training
RF_GK <- randomForest(AverageRatingOverall ~ ., data = TrainGKRF)

#Rank Prediction
TestGkRF$PredictedRanking <- predict(RF_GK, newdata = TestGkRF)
InputNorGK$PredictedRankingRF <- predict(RF_GK, newdata = InputNorGK)

RatingsGkKPIRF <- data.frame(InputNorGK$FullName,round(InputNorGK$PredictedRankingRF,2))

#MAE Calculation
MAEGkKpiRF <- mean(abs(TestGkRF$AverageRatingOverall - TestGkRF$PredictedRanking))
MAEGkKpiRF
```

8. Appendix B- Ethics Form

Internal research ethics application form

For module LUBS5579M (Business Analytics and Decision Sciences Dissertation) covered by University of Leeds ethical approval reference AREA 17-055[1].

Student ID	201678076
Student Name	Abhishek .
Degree Programme	MSc Business Analytics and Decision Sciences
Title / topic area	Quantitative Analysis of Footballer Performance in the English Premier League using Machine Learning Algorithms.
Name of dissertation supervisor	Dr. Sajid Siraj

Are you planning to work with (data on) human participants for your dissertation?	Please tick the relevant box
Yes (This includes interviews, surveys and secondary data analysis of social media or internet data).	
No, I am conducting an in-depth literature review with analysis.	✓

If you ticked 'No' you do not need to take further action in respect of ethical approval. Please proceed to the declarations on page 7.

If you ticked 'Yes' you need to complete the rest of this form.

You MUST discuss your research design and the ethical issues it raises with your dissertation supervisor and receive their signed approval before you approach any participants or collect any data.

You MUST include a copy of your ethics form (signed by your supervisor) as an appendix to your final dissertation submission (both the electronic and paper copies).

[1] Ethical approval valid until 31/12/2027.

INTERNAL RESEARCH ETHICS APPLICATION
Part A: Compliance with the module's block ethical approval

Ethical review is required for **all research involving human participants**. Further details of the University of Leeds ethical review requirements are provided in the *Research Ethics Policy* available at: http://ris.leeds.ac.uk/ResearchEthicsPolicies and www.leeds.ac.uk/ethics.

1. Will your dissertation involve any of the following?	**Yes**	**No**
New data collected by administering interviews for analysis		
New data collected by administering questionnaires for analysis		
New data collected from observing individuals or populations		
Working with secondary aggregated or population data		
Using open data		
Using data given to you by a company		
Any other research methodology, please specify:		

2. Will any of the participants be from any of the following groups? (Tick as appropriate)	**Yes**	**No**
Children under 16		
Adults with learning disabilities		
Adults with other forms of mental incapacity or mental illness		
Adults in emergency situations		
Prisoners or young offenders		
Those who could be considered to have a particularly dependent relationship with the investigator, e.g. members of staff, students		
Other vulnerable groups, please specify:		

3. Will the project/dissertation/fieldwork involve any of the following: (You may select more than one)	**Yes**	**No**

Patients and users of the NHS (including NHS patients treated under contracts with private sector)		
Individuals identified as potential participants because of their status as relatives or carers of patients and users of the NHS		
The use of, or potential access to, NHS premises or facilities		
NHS staff - recruited as potential research participants by virtue of their professional role		
A prison or a young offender institution in England and Wales (and is health related)		

If you have answered 'Yes' to **ANY of the above statements in questions 2 or 3** then you will need to apply for full ethical approval which is a faculty committee level process. This can take up to 6-8 weeks, so it is important that you consult further with your supervisor and/or program director for guidance with this application as soon as possible. Please now complete and sign the final page of this document. The application form for full ethical review and further information about the process are available at http://ris.leeds.ac.uk/uolethicsapplication.

If you answered **'No' to ALL of the statements in Questions 2 and 3** please continue to part B.

INTERNAL RESEARCH ETHICS APPLICATION
Part B: Ethical considerations within block ethical approval

4. Will the research touch on sensitive topics or raise other challenges?	Yes	No
Will the study require the cooperation of a gatekeeper for initial access to groups or individuals who are taking part in the study (eg students at school, members of self-help groups, residents of a nursing home)?		
Will participants be taking part in the research without their knowledge and consent (eg covert observation of people in non-public places)?		
Will the study involve discussion of sensitive topics (eg sexual activity, drug use)?		
Could the study induce psychological stress or anxiety or cause harm or have negative consequences beyond the risks encountered in normal life?		
Are there any potential conflicts of interest?		
Does any relationship exist between the researcher(s) and the participant(s), other than that required by the activities associated with the project (e.g., fellow students, staff, etc)?		
Does the research involve any risks to the researchers themselves, or individuals not directly involved in the research?		

If you have answered 'Yes' to any of the statements above please describe the ethical issues raised and your plans to resolve them on a separate page. Agree this with your supervisor and submit it with this form.
You MAY be referred for light touch or full ethical review.

5. International Research Does your research involve participants outside of the UK?	Yes	No
Are any of your research participants located outside of the UK? For example: will you be gathering data through Skype interviews with participants located overseas?		
Will any of the fieldwork or research require you to travel outside of the UK to collect data?		

If you have answered 'Yes' to any of the statements above please describe the ethical issues raised with: gaining consent and gathering data from participants located overseas, securely storing and transferring data from the field back to the UK, any cultural issues that may be relevant. Please outline your plans to resolve this on a separate page and ensure that you have completed a risk assessment form (available from LUBS student education). Agree this with your supervisor and submit it with this form.
You MAY be referred for a light touch or full ethical review if you are unable to demonstrate that you have resolved the ethical issues relating to international research.

6. Personal safety Where will any fieldwork/ interviews/ focus groups take place?	Yes	No
At the university or other public place (please specify below).		
At my home address		
At the research subject's home address		

Some other location (please specify below).		

If you conduct fieldwork anywhere except at the university or other public place you need to review security issues with your supervisor and have them confirmed by the Module Leader who may refer you for a light touch or full ethical review. Write a brief statement indicating any security/personal safety issues arising for you and/or for your participants, explaining how these will be managed. Agree this with your supervisor and submit it with this form. Please note that conducting fieldwork at the research subject's home address will require strong justification and is generally not encouraged.

7. Anonymity	**Yes**	**No**
Is there any potential for data to be traced back to individuals or organisations, for instance because it has been anonymised in such a way that there remains risk? for example: highlighting people's positions within an organisation, which may reveal them		

If you have answered 'Yes' to question 7, please discuss this further with your supervisor. You need to provide a strong justification for this decision on a separate sheet. This application will need to be reviewed by the dissertation Module Leader and may require a full ethical review.

8. Data management issues

Will the research involve any of the following activities at any stage (including identification of potential research participants)?	**Yes**	**No**
a. Examination of personal records by those who would not normally have access		
b. Sharing data with other organisations		
c. Use of personal addresses, postcodes, faxes, e-mails or telephone numbers		
d. Publication of direct quotations from respondents		
e. Publication of data that might allow identification of individuals to be identified		
f. Use of audio/visual recording devices		
g. Storage of personal data on any of the following:		
FLASH memory or other portable storage devices (e.g. USB storage)		
Home or other personal computers		

	Private company computers		
	Laptop computers		

If you have answered 'Yes' to any of the questions above you must ensure that you follow the University of Leeds Information Protection Policy: http://www.leeds.ac.uk/informationsecurity and the Research Data Management Policy: http://library.leeds.ac.uk/research-data-policies#activate-tab1_university_research_data_policy.

You are obliged to provide a copy of your anonymised data to your supervisor for their records and to destroy other copies of your data when your degree has been confirmed.

Dissertation Research Ethical Approval: Declaration

For students	*Please tick as appropriate*
Option 1: I **will NOT** conduct fieldwork with (data on) human participants for my dissertation.	
Option 2: I **will** conduct fieldwork with (data on) human participants for my dissertation.	

For **options 1 and 2 -** I confirm that:

- The research ethics form is accurate to the best of my knowledge.
- I have consulted the University of Leeds Research Ethics Policy available at http://ris.leeds.ac.uk/ResearchEthicsPolicies.
- I understand that ethical approval will only apply to the project I have outlined in this application and that I will need to re-apply, should my plans change substantially.

For **option 2** only:

- I am aware of the University of Leeds protocols for ethical research, in particular in respect to protocols on **informed consent**, **verbal consent**, **reimbursement for participants and low risk observation**. If any are applicable to me, signing this form confirms that I will carry out my work in accordance with them. http://ris.leeds.ac.uk/PlanningResearch

Student's signature: ……Abhishek ……..……..

Date: ……5th August 2023
…….……..……..

For supervisors	**Yes**	**No**
No further action required		
I confirm that the dissertation is in line with the module's block ethical approval (Part A & question 8).	✓	
I have discussed the ethical issues arising from the research with the student and agree that these have been accurately and fully addressed.	✓	
I have reviewed the student's research proposal.	✓	
I have reviewed the student's Risk Assessment Form (*if necessary*).		
Further actions required		
Refer to dissertation Module Leader for further review / discussion.		
The dissertation falls outside the module's block ethical approval and the student was advised to apply for full ethical review.		

Supervisor's signature: ……S.S…….……..……..

Date: ……5th August 2023 …….……..