National Technical University of Athens
School of Applied Mathematics and Physics

Predicting the Future Performance of Soccer Players

Master Thesis

of

EVANGELOS A. GKIASTAS

Supervisor: Dimitris Karlis
Professor

Athens, July 2022

National Technical University of Athens
School of Applied Mathematics and Physics

Predicting the Future Performance of Soccer Players

Master Thesis

of

EVANGELOS A. GKIASTAS

Supervisor: Dimitris Karlis
Professor

Approved by the examination committee on 20th July 2022.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Dimitris Karlis	Ioannis Ntzoufras	Dimitris Fouskakis
Professor	Professor	Professor

Athens, July 2022

National Technical University of Athens
School of Applied Mathematics and Physics

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this master thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Evangelos A. Gkiastas

20th July 2022

Abstract

Soccer is one of the most widespread and popular sports on the planet and the increasing amount and utilization of data in all aspects of life, could not leave, of course, this field unaffected. The area of sports analytics has attracted a lot of interest in recent years, with applications that affect many aspects of the game. While match outcome predictions, injury preventions, team tactics improvement and betting odds estimations have been widely investigated with different data-driven approaches, future player performance prediction is quite unexplored issue. A perspective of how the athlete's performance will develop in the near or distant future can significantly impact both the athlete and the team on a variety of different levels. This thesis focuses particularly on this issue and is divided in three parts. Firstly, a real-world dataset of elite soccer player games is collected and created, containing individual and team attributes, as well as pre game and other information related to performance. Secondly, it is investigated which variables appear to be more highly associated with the prediction of player's rating in a future game. Finally, the third contribution of this thesis is the forecasting of individual performance in specific future games. Various statistical, machine learning and deep learning models are applied to multivariate time series data with success, producing much better results compared to those of random and naive predictors.

Keywords

Soccer, sports analytics, data mining, machine learning, deep learning, time series forecasting, player performance prediction

Περίληψη

Το ποδόσφαιρο είναι ένα από τα πιο διαδεδομένα και δημοφιλή αθλήματα στον πλανήτη και η αυξανόμενη ποσότητα και αξιοποίηση των δεδομένων σε όλες τις πτυχές της ζωής, δεν θα μπορούσε να αφήσει, φυσικά, αυτό το πεδίο ανεπηρέαστο. Ο τομέας των αθλητικών αναλύσεων έχει προσελκύσει μεγάλο ενδιαφέρον τα τελευταία χρόνια, με εφαρμογές που επηρεάζουν πολλές πτυχές του παιχνιδιού. Ενώ οι προβλέψεις αποτελεσμάτων αγώνων, η πρόληψη τραυματισμών, η βελτίωση της τακτικής των ομάδων και οι εκτιμήσεις στοιχηματικών αποδόσεων έχουν διερευνηθεί ευρέως με προσεγγίσεις βάσει δεδομένων, η μελλοντική πρόβλεψη της απόδοσης ενός ποδοσφαιριστή είναι ένα αρκετά ανεξερεύνητο θέμα. Η γνώση του πώς θα εξελιχθεί η απόδοση του αθλητή στο εγγύς ή μακρινό μέλλον μπορεί να επηρεάσει σημαντικά τόσο τον αθλητή όσο και την ομάδα σε διάφορα επίπεδα. Η παρούσα διπλωματική εργασία εστιάζει ιδιαίτερα σε αυτό το θέμα και χωρίζεται σε τρία μέρη. Πρώτον, ένα σύνολο πραγματικών δεδομένων από παιχνίδια κορυφαίων ποδοσφαιριστών συλλέγεται και δημιουργείται από το μηδέν, περιέχοντας τα ατομικά χαρακτηριστικά, στοιχεία για τις ομάδες και το παιχνίδι, καθώς και άλλες πληροφορίες που σχετίζονται με την απόδοση. Επιπλέον, ερευνάται ποιες μεταβλητές φαίνεται να συνδέονται περισσότερο με τη διερεύνηση της επίδοσης του παίκτη σε μελλοντικό παιχνίδι. Τέλος, η τρίτη συνεισφορά της παρούσας εργασίας είναι η πρόβλεψη της ατομικής απόδοσης των ποδοσφαιριστών σε συγκεκριμένα μελλοντικά παιχνίδια. Διάφορες στατιστικές μέθοδοι, μοντέλα μηχανικής μάθησης και βαθιάς μάθησης εφαρμόζονται σε χρονοσειρές πολλών μεταβλητών με επιτυχία, παράγοντας πολύ καλύτερα αποτελέσματα σε σύγκριση με αυτά που δίνουν τυχαία και απλοϊκά μοντέλα.

Λέξεις Κλειδιά

Ποδόσφαιρο, αθλητικές αναλύσεις, εξόρυξη δεδομένων, μηχανική μάθηση, βαθιά μάθηση, πρόβλεψη χρονοσειρών, πρόβλεψη απόδοσης παικτών

Acknowledgements

First of all, I would like to thank my primary supervisor Dimitrios Karlis. His knowledge, motivation and guidance throughout this thesis, were quite helpful.

I would also like to thank my parents, family and friends for their constant support and encouragement.

Athens, July 2022

Evangelos A. Gkiastas

Table of Contents

List of Figures

List of Images

List of Tables

Chapter 1

Introduction

In recent years, the field of sports analytics has begun to attract a lot of interest. Predictive modeling is one of the methods for quantitative research in sports that can be used because of the increase in data volume, computing power, and the creation of effective algorithms. This differs from the conventional approach to studying sports, which employs control groups and significance testing. For a modern sports professional, predictive modeling can be a useful and effective tool because it enables them to predict outcomes directly rather than relying on educated guesses.

Specifically in soccer, we see top tier clubs to hire data experts and initiate data science departments [1] [2]. Event data, optical pitch view for technical analysis, expected goals metric and more terms have been introduced in our life for some years now. Many soccer clubs use data in some way to help them make decisions, but how and how much they use it varies greatly from club to club. Manchester City's midfielder, Kevin De Bruyne, attracted media attention recently, because he negotiated his new contract, providing a analytical report of his past and future importance to the team, in collaboration with a data analytics company [3].

Concerning future individual performance, remains an untapped field. While there have been several attempts for predicting sport games' outcomes using data from the past, or infer player ratings using post match statistics, there are almost no attempts to predict the performance of soccer players on future games. Usage of predictive modeling techniques can indicate a way to explore the potential of this idea.

1.1 Motivation

"O jogo bonito" is the Portuguese term for the "beautiful game". There is no agreed-upon reason why soccer is called the "beautiful game", but many people love it because it is so unpredictable, has the power to unite communities through fandom, and can be aesthetically pleasing when played well. Soccer greats like George Best, Pele, Diego Maradona, Ronaldinho, Cristiano Ronaldo and Lionel Messi are skilled masters who only serve to enhance the sport's beauty. Its appeal is also derived from the spectacular goals that are scored using volleys, chips, lobs, and headers.

[1] https://www.bbc.com/news/business-56164159
[2] https://www.hudl.com/blog/how-monchi-and-sevilla-fc-use-data-in-scouting
[3] https://www.mirror.co.uk/sport/football/news/kevin-de-bruyne-uses-data-23870686

These legends along with numerous super stars are idols for millions of people in the world, that wait for their next great appearance, read the news about their lifestyle and buy their shirts. This every day interaction with soccer brought the need to know how a footballer will perform and is the main motivation behind this thesis.

1.2 Problem formulation

To date, there aren't many methods that try to analyze past data and infer the performance of the player. Any researcher interested in this field should spend a lot of time collecting data, and even more to create valuable explanatory input. Therefore the problem that this dissertation focuses on is:

Can statistical and machine/deep learning methods predict successfully the future performance of a player in a soccer game?

The problem can be divided into the following sub-questions:

1. Is it possible to create a dataset of soccer players and their games from scratch?
2. Which attributes are important when focusing on individual performance?
3. Can future player performance be predicted using multivariate time-series forecasting models?

1.3 Constraints

There is a variety of data that can be used to answer this problem. However, the availability of them is debatable. The thesis only uses data from a fixed set of players and all these players will be treated the same, even though in real life they have different history. Most importantly, performance of a player is a complex topic that correlates with non recorded factors, like personal life events and mood, which cannot be taken into account on this research.

1.4 Structure of this thesis

The rest of the thesis is written out as follows.

Before moving on to player performance, Chapter 2 gives a general overview of the nature of sports analytics and related fields. It begins with a review of previous studies on the state of academic sports analytics research and available commercial solutions. After that, it presents the most recent work in sports analytics for performance, highlighting the research gap and laying out the motivation for this thesis.

The statistical and machine learning techniques that were applied throughout this study, as well as the evaluation metrics and the software aspects, are described in Chapter 3.

In Chapter 4, some of the problems while working with data are introduced, along with their relevance to soccer research in this thesis. The method for collecting and preprocessing data is then presented in detail.

Chapter 5 tries to interpret which match or player features are significant and affect the performance of the athlete. Regression techniques are also used to predict a player's performance in a soccer match.

The other investigation into using time series forecasting to predict a player's performance is covered in Chapter 6. This chapter starts with a more detailed look on the input creation, the algorithms' design components and provides detailed comparison of the results in correlation with naive models.

Chapter 2

Background Theory

This chapter aims at describing the theory needed to understand the work conducted, as well as present related work that is useful for this thesis. It starts with an introduction to the field of sports analytics and continues with the soccer player performance research. Finally, some basic ideas about data mining and machine learning are presented.

2.1 Sports Analytics Landscape

Sports analytics refers to the use of historical data from the past and sophisticated statistics to evaluate performance, make data-driven choices, and forecast future outcomes in sports. Giving a team or person a competitive edge is the goal of that use. Sports analytics focuses on two primary areas. analytics both on and off the field. The goal of on-field analytics is to help teams and players perform better. It takes care of things like player fitness and game strategy. The commercial aspect of sports is covered by off-field analytics. Off-field analytics is focused on helping a sports team or organization uncover patterns and insights in data that might increase ticket and product sales, enhance fan interaction, etc.

Data collection has gotten more thorough and achievable with the advancement of technology over the past several years. Sports-specific technologies that enable things like game simulations by teams prior to play, improved fan acquisition and marketing strategies, and even understanding the impact of sponsorship on each team have all been made possible by advances in data collection, which has allowed sports analytics to develop as well.

Sports gambling is another important area where sports analytics have had a big influence on professional sports. In-depth sports analytics have raised the bar for sports betting, enabling players to make better decisions whether they are participating in nightly wagers or fantasy sports tournaments. Numerous businesses and websites have been created to aid in giving fans the most recent information for their betting demands.

2.1.1 Literature overview

Despite the fact that the phrase "sports analytics" is relatively recent, research that falls under the present definition of sports analytics has been conducted over the past few

decades. According to Wright (2009) [55], operations research has been used in sports for more than 50 years. Since the middle of the 20th century, we see that predicting soccer scores has been a significant study topic. Moroney (1956) [41] and Reep (1971) [47] employed the Poisson distribution and negative binomial distribution to estimate the number of goals scored in a soccer match based on previous team outcomes.

From then until now, there has been great progress in research in numerous sports, that doesn't only focus on predicting match outcomes. Rajiv Shah (2016) [51] used sequence modeling and a dataset of over 20,000 three pointers from NBA to learn the trajectory of a basketball, with recurrent neural networks and determined the likelihood that a three-point attempt will be successful. Lee Jae Sik (2022) [30] created a system for predicting baseball pitches -the spot where the pitched ball lands among the fictitious grids- using ensemble models of deep neural networks. Bartolucci and Murphy (2015) [3] developed a 24-hour ultramarathon performance and strategy analysis using a finite mixture latent trajectory model and grouped runners according to their speed and inclination to take breaks, facilitating clustering techniques. Sharma et al (2017) [52] gathered data from the sensor and the camera to examine a tennis player's swing consistency, the point that makes the biggest and most significant difference, and the areas that athlete should concentrate on to enhance their performance.

Although the popularity of sports analytics is constantly increasing, the field is still considered as fragmented. Many writers stop working on their projects after they publish them and the area is viewed as being too limited to support specialized university programs. Additionally, many findings might not be published, because the relevant sports professionals who participated in a study might want to use them to gain an advantage over rivals. The fact that there are still few conferences and journals that are appropriate venues for sports analytics research is another issue. Many studies that fall under the category of "sports analytics" have been published in non-sports analytics journals, like economic ones or journals in computer science.

Despite the field's fragmentation, a clear upward trend in popularity can be seen. From that vantage point, it is evident that sports analytics is a thriving area of study. It is not difficult to notice that there are some patterns on which research is increasingly active, even though many of the papers that are published each year are standalone. In the next subsection we explore these trends, specifically for soccer.

2.1.2 Soccer research

Following is a breakdown of a few of these trends:

Predicting the results of games or competitions: A large percentage of research is devoted to this topic. There are various efforts to approach this issue:

- Dixon and Coles (1997) [13] created a model that followed the Poisson distribution to provide probability for game outcomes and scores. The Dixon and Coles approach gained a lot of traction and served as a standard for future models. The predicted number of goals for each side were translated to match result probabilities using a Poisson regression model. Model included parameters related to past performance,

along with some realistic refinements.

- Goddard (2005) [20] developed a regression model that takes into account recent form, team ability, game importance, and geographic distance and looked for team rivalry using the variable of geographic distance, creating one of the earliest publications that took into account factors other than actual game results. He compared his simulation results with the betting markets, which allowed him to determine the chances of a positive gambling return, concluding that algorithms based on outcomes and methods based on scores have similar levels of accuracy and that a hybrid model would be the most effective approach.

- Groll and Abedieh (2013) [21] used data from the EURO 2004 and 2008 to analyze the impact of various co-variates on the success of national sides in terms of the number of goals they score in single matches of EURO 2012. This was done using a pairwise generalized linear mixed Poisson model for the number of goals scored by national teams facing off in European soccer championship matches. The main goal was to examine the explanatory power of bookmakers' odds in this situation and, by incorporating more attributes, to gain insight into which co-variates might provide information beyond that provided by odds and, second, which attributes have already been covered by bookmakers' odds.

- Elmiligi and Saad (2022) [14] introduced a novel hybrid approach, in order to forecast the results of upcoming soccer matches, combining statistical models and machine learning. The research examined the hidden patterns in a training dataset that contains the results of two hundred thousands soccer matches played between the 2000/2001 and 2016/2017 seasons. They used feature engineering techniques to investigate individual leagues and team statistics, discussing the effect of playing at home or away on winning the match, comparing the efficacy of using only recent match outcomes data to all matches in the training set, and evaluating the prediction accuracy when creating separate models for each league versus a single model for all leagues.

Examining movements and tactics: Some research focuses on analyzing similar factors in sports on a personal or group level. Several instances include:

- Haase and Brefeld (2013) [23] investigated the challenge of quickly identifying related movements in positional data streams given a query trajectory. The strategy was based on a representation of movements that is translation-, rotation-, and scale-invariant and then a query trajectory was used to efficiently compute nearby neighbors via dynamic temporal warping and locality-sensitive hashing. Using positional data streams captured from an actual soccer match, they found players with frequently similar movements, which can be very effective in team tactics.

- A player's positioning throughout each second of the game, the overall distance they cover, their "preferred" actions, and the ball were all studied using computer vision by Stein et al. (2018) [53], who concentrated on data gathered only through

video analysis in soccer. To make data collection and editing easier, the stadium's three-dimensional image was first downsized to two dimensions. The performance of a team was then inferred by using particular algorithms. The most well-known businesses in this industry employ this technology, which demonstrates its great success. The study came to the conclusion that it is a useful strategy that will be applied in the future, while mentioning some disadvantages and constraints.

- Stockl et al. (2021) [39] put out a brand-new Graph Neural Network architecture that can handle unstructured data and can immediately learn from a simple feature representation of the team behavior of the soccer players on the field. They developed a tool-set to assess the impact of defensive strategy on the opponent at team and player levels, based on the model outputs. Disruption Maps produce a concise visual depiction of a team's impact on the xThreat ("probability of a shot occurring right after a pass to an attacker"), xPass ("likelihood of a successful pass to an attacker at any moment"), and xReceiver ("probability of every player to receive a pass at any moment of possession") values of the opposition relative to their global overage, enabling users to see areas where a defense has excelled or suffered based on lowering/raising an opponent's threat, pass danger, and player availability. This study made possible to assess defensive behavior and offers coaches and supporters useful resources and information.

Analyzing and forecasting injuries: Soccer injuries have been measured in a variety of ways in academia.While some of these tests monitor an athlete's overall fitness, several researches have concentrated on certain risk factors.

- Henderson et el. (2010) [25] used logistic regression to check how several physical and performance factors, including anthropometry, flexibility, lower limb strength and power, speed, and agility, affect the occurrence of hamstring injuries in a group of English Premier League soccer players. All hamstring injuries over the following 45-week competitive season were identified and documented. Age, lean mass, jump performance, and active hip flexion range of motion were features that were noticeably correlated with greater propensity for hamstring injury. Multiple logistic regression methods used to link individual physical and performance capabilities with tendency to sustain a hamstring injury.

- In order to aid in the diagnosis and prognosis of anterior cruciate ligament reconstructed (ACL-R) individuals throughout recovery, Arosha Senanayke et al. (2014) [2] created a knowledge-based framework employing hybrid intelligence approaches, and verified it for a sample of healthy and ACL-R subjects with extremely high retrieval accuracy. With the help of clustering integrated kinematics and EMG data, they created an example repository of healthy and varied ACL-R cases at various phases of recovery for the CBR model. The classification model based on FURIA worked well for all activities, according to the results and gave doctors and physiatrists useful feedback in assessing the participants' progress in rehabilitation following ACL injury or surgery, showing that a complete system for ACL injury avoidance

and sports performance assessment can be provided by keeping the players' pre-injury, post-injury, and post-surgery profiles.

- Rossi et al. (2018) [49] developed an automated, machine learning-based multi-dimensional approach to injury prediction in soccer, demonstrating that the injury forecaster can offer a balance between accuracy and interpretability, minimizing the frequency of false alarms in comparison to cutting-edge alternatives, while still offering a straightforward set of guidelines to help comprehend the causes of the observed injuries. They also showed that their method can be utilized early in the season enabling teams to save a substantial portion of the expenditures associated with seasonal injuries. This mechanism assessed and analyzed the intricate relationships between injury risk and training effectiveness in professional soccer.

2.1.3 Commercial applications

The market of sports-related services and products by organizations, single proprietorships, and partnerships that provide spectator sports and participating sports make up the sports market. In 2020, the value of the world sports market was close to $388.3 billion. With a drop mostly brought on by lockdown and social segregation restrictions imposed by many nations, as well as a global economic slump brought on by the COVID-19 breakout and the containment efforts. By 2025 and 2030, respectively, the market for sports is projected to reach $599.9 billion and $826.0 billion. Rapid urbanization, the expansion of emerging economies, and the introduction of several channels to attract viewers all contributed to growth over the historical period. Future development will be fueled by the emergence of e-sports, increasing sponsorships, and an increase in internet-enabled gadgets. [1]

Sports analytics market is expected to grow from USD 2.1 billion in 2020 to USD 16.5 billion by 2030. The primary factor driving the market expansion is the rising inclination of team managers and coaches for adopting cutting-edge technology to use real-time data to create a variety of game tactics and schedule training sessions. Additionally, smart sports technology is being adopted more widely throughout the world to get quantitative data for enhancing game performance. The overall demand for the sports analytics market is also predicted to expand with the rise in investment by sports organizations in adopting a data-driven decision-making approach, from player recruiting to fan engagement. [2]

The breadth of sports analytics has grown as a result of recent technological advancements in data collecting and administration. In most major sports, the use of data, big data, and statistics has increased significantly. Companies change throughout time as a result of changes in the market or the introduction of new technologies. As a result of the emergence of new businesses and new trends as a result of the development of contemporary technology, a list of the top players in sport analytics, and specifically in soccer is provided below:

[1] https://www.thebusinessresearchcompany.com/report/sports-market
[2] https://www.globenewswire.com/en/news-release/2022/05/27/2452062/0/en/Sports-Analytics-Market-a-16-5-billion-Industry-by-2030-with-a-CAGR-of-22-9-Strategic-Market-Research.html

- Opta Sports[3], which was founded in 1996, has been the industry leader in sports data distribution and gathering for more than 20 years, by providing media customers and clubs statistical data and player performance statistics from the world's top sport leagues. Their most well-liked statistic to date has been xG (expected goals), which assigns a value to a particular shot, or set of shots, to assess the chance that the shot will be successful based on previous information about comparable shoots. The areas of sequences and defensive coverage, however, where they have devised a metric to assess the area of defensive activities by a player throughout a match, represent their most recent advancements. They are now part of StatsPerform, another pioneer in sports data collection and predictive analysis for use across various sports sectors.

- STATSports[4] is a well-known manufacturer of GPS monitoring vests, was founded in Ireland and now works with prestigious teams including Manchester United, Juventus, Manchester City, the Carolina Panthers, and the New York Knicks. Provides equipment that athletes may wear throughout practices and games to collect detailed information on their physical condition and performance. The landscape of GPS player monitoring devices has changed as a result, enabling managers and coaches to decide how to modify individual or team workloads depending on the information supplied. They also offer non-professional athletes an easier bluetooth-enabled GPS tracking app in addition to that. The 14 metrics available to athletes include speed and distance, and they may compare their performance across leagues and leaderboards.

- SportsRadar[5] is a sports technology firm that is well-positioned at the intersection of the media, betting, and sports industries. It offers a variety of solutions to sports federations, news outlets, consumer platforms, and sports betting operators to assist them expand their businesses. Sports leagues including the NBA, NHL, MLB, NASCAR, FIFA, and UEFA are among their partners. They employ specialists who conceive, create, and implement predictive models and algorithms, increasing and automating sportsbook risk management services and liquidity-driven odds trading. To solve real-world issues with data collecting and production for Sports, they also use machine vision and deep learning.

It goes without saying that the current sports analytics tools on the market may assist the coaching staff in making more informed judgments by offering a wealth of facts that might not have otherwise been available in an accessible format, such as graphs and tables. However, the software that is now on the market ignores more sophisticated capabilities that can be provided by more sophisticated machine learning approaches and can aid in an athlete's or a team's decision-making process.

[3] www.statsperform.com/opta/
[4] https://statsports.com/
[5] https://sportradar.com/

2.2 Player Performance Prediction Research

Predicting how well or bad will a player perform in the next match is not something that is widely done today, like the match outcomes for example, that is done by both soccer experts and machine learning algorithms, with various results. This section discusses some challenges in predicting the performance of a player and some previous work.

2.2.1 Challenges of predicting future individual performance

Predictions made by humans and computers both have their limitations. People have emotions, which can influence how teams are analyzed before matches and, consequently, how predictions are made. The psychological condition of the squad at the moment is unknown to a computer. Interpersonal conflicts between the players and the coaches can have an impact. Determining which traits are crucial is an issue that both people and computers face.

Soccer is very stochastic, just as many other sports. Among hundreds of shots, passes, and dribbles, one fortunate hit can ultimately alter the entire result of the game. This makes it more difficult for both humans and computers to forecast the results of soccer matches.

2.2.2 Factors that affect a soccer player

Although rating systems, such as the one we will examine in more detail in the data gathering sections, can be used to assess player skill, there are still a number of considerations. The relationships between teammates is a crucial one. It often takes into account each player's talents and shortcomings as well as the team's culture and mentality, and it seems to be a crucial indicator of how well a team performs. Player relationships and more fundamental things are described in detail by Gréhaigne et al. (2005) [22].

The team fitness is possibly the most popular performance metric in sport and may affect an individual performance. Experts agree with this figure, and numerous soccer statistics websites have embraced it. There are also articles that reveal peer effects on individual performance in a team sport like Molodchik et al. (2021) [40].

The psychology of the players can also have an impact on how well they perform. This parameter cannot be measured. To compare the psychological makeup of individuals playing in the same area, Constantinou et al. (2012) [9] used information on motivation to win and team spirit.

2.2.3 Related work

To date, many of the published works mentions "player performance prediction", although none of them tries to forecast if the next game of a player will be good or bad, like they do with injury events. Artificial intelligence methods were used for "match attendance," "technical and tactical analysis," and "psychological dynamics of cooperating" like Claudino et al. (2019) [8]. Most of the research focuses on performance as a rating system and suggests methods of rating a player using post game statistics. Also, some of

them like Matteo et al. (2022) [37] and Porter (2018) [45] try to forecast fantasy points of a player.

Below we list some of the different approaches to this issue:

- Hughes and Bartlett (2010) [27] and Lames and McGarry (2007) [35] describe sports performance in terms of one or more performance indicators, which are groups of action variables. Some of them are shots, ball possessions, combinations and more and they are used by coaches and performance analysts to discuss or compare the good and bad elements of performance within or between contests.

- Brooks et al. (2016) [6] provided a player rating scheme that is only dependent on the quality of passes made. This number was determined from the correlation between pass locations during a possession and the creation of shooting possibilities, by using event data from the 2012-2013 La Liga season. A supervised learning model has been applied to understand this association and then each pass was assigned a weight based on how important it is for setting up a shot later in the possession, producing a data-driven ranking of the players.

- Pappalardo et al. (2019) [43] introduced PlayeRank, a system based on data that provides a multi-dimensional and role-aware assessment of soccer players' performance. The experimental analysis had a sizable collection of soccer-logs—18 tournaments, 31 million events, and 21 thousand players and revealed a number of intriguing characteristics that describe soccer players' performance. This tool can be used in order to evaluate, search for, rank, and propose soccer players, by professional soccer scouts.

- Pariath et al. (2018) [44] attempted to predict performance not in the future, despite the title's name, but just as dependent variable along with other factors like different positions in the game and skills extracted from the EA Sports FIFA video game. They provided a linear regression model that establishes a correlation between the soccer players' performance characteristics and their total performance value, trying to aid in the scouting and coaching of soccer players. They came to the conclusion that the player's total performance appears to be a better quality to consider for recruiting since the market value may be altered by players' release clauses with the club, age, league competition, or club budget.

- Like the previous publication, Pantzalis et al. (2020) [42] mention player performance prediction, but they discuss it from the scope of a rating system and not in the future. The goal of this study was to determine which traits and data are most important in defining a central defender as a top-tier player by comparing them to a central defender's season rating. They used player attributes, playing positions and some demographic features in their methods to evaluate a player.

- Shultze and Wellbrock (2017) [50] and Kharrat et al. (2020) [31] used plus/minus metric for individual soccer player performance. This metric, in its simplest form checks player's implied effect on his team's goal difference while he is on the field of

play. Both approaches provided a quick way to evaluate player performance above the traditional descriptive statistics and human video analysis.

- Arndt and Brefeld (2016) [1] provided regression-based methods, to forecast future soccer player performances. Players-specific models could be learned with multitask generalizations of ridge regression and support vector regression methods. They also introduced a modified recursive feature removal technique because the optimization required a large number of features. Their findings using actual data from the German Bundesliga demonstrated that the amount of data is frequently insufficient to learn individual player models.

As we've seen, there isn't much literature in this field. Additionally, the research cited above, only ranks individual performance, using post game statistics. To the author's knowledge, there hasn't been any academic study, except the last one above, on future performance prediction of soccer players. By separating the complex correlations between the various elements that go into a successful or unsuccessful game, predictive modeling in soccer (or other sports) player performance would significantly advance the subject of sports science.

2.3 General Terms

This section of the dissertation defines a few general words related to the subject that will be examined. To have a broader perspective on the matter, it is crucial to draw attention to them.

2.3.1 Data mining

In the 1980s, as the computer era dawned and more data began to be created, there was developed a demand for more automated techniques of data analysis. The database language Structured query language, better known as SQL, was also introduced in 1980. It quickly shot to popularity and increased the quantity of data that could be stored. The databases grew so large as more businesses began to use computers in their administration to store, for example, financial records that the traditional methods, where they performed statistical analysis by hand, couldn't keep up with the new requirements of efficiency, necessitating the development of new techniques, Hand (2007) [24].

This prompted businesses to combine their research funding, which resulted in significant developments in fields like artificial intelligence, neural networks, machine learning, and statistics. A press book titled Information Discovery in Databases, published at the start of 1990 by the organization for the advancement of artificial intelligence, detailed how to find hidden knowledge in data, Frawley et al. (1992) [15]. Data mining, which has become a buzzword for the whole field of data analysis, may be thought of as a more advanced form of the knowledge discovery in databases procedure described in that book's analysis stage.

Data mining is described as a method for obtaining useful information from a bigger collection of any raw data. It suggests utilizing one or more pieces of software to analyze

data patterns in massive data sets. It may be thought of as a place where computer science, machine learning, and statistics meet. Regression, association learning, data classification, and clustering are some methods that fall within this group Hand (2007) [24].

In this point it's important to highlight the term of big data. While until a few years ago this concept was barely known, today is one of the most talked about topics in the whole business people. Big Data is defined as huge amounts of data, structured semi-structured or unstructured, with the ability to extract information from them, where due to volume no longer matters a single record but their total size. The main concepts that form the basis of the definition of Big Data consist of the model of three V:

- Volume: Vast volumes of data in the terabyte and zettabyte ranges.
- Velocity: Massive volumes of data from fast transactions generate partition streams quickly, cutting down on the amount of processing time that is available to them. Real-time streams processing has replaced the previous method of processing static data.
- Variety: Data are gathered from many sources. They can take on a number of shapes, including:
  - Structured: Excel sheets, SQL tables, etc.
  - Semi-structured data include XML documents like tweets and emails.
  - Text, pictures, videos, etc. that are unstructured

Big Data is described by De Mauro et al. (2016) [11] as high-volume, high-speed, high-variety data that require novel processing models with enhanced knowledge, automation, and decision-making. It is increasingly recognized that the term "big" refers to more than just volume and that, given how quickly technology develops, the big data of today may not be the big data of tomorrow. Some examples of data retrieved by all organizations today are:

- Web data (user navigation, cookies)
- the text data (Facebook posts, news, email)
- time and location data via GPS, WI-FI, smartphones etc
- data from sensors and intelligent networks (from cars, oil pipelines, etc)
- social network data, such as Facebook, Instagram etc

All of these are used either to extract knowledge from personal data and user interests and then commercial exploitation, or to improve the performance of all the technologies mentioned.

2.3.2 Machine learning

In recent years, both in academic circles and in business, the subfield of artificial intelligence known as machine learning—particularly neural networks and deep learning—has grown in prominence. Machine learning aims to utilize data to learn to make predictions or judgments, as opposed to traditional rule-based artificial intelligence, where an algorithm is just a list of static rules that have been predefined. Due to the complexity or number of restrictions in many situations, it is often impossible to develop or design rules to aid in problem solving. In these cases, machine learning can use data to discover a solution, Mitchell (2006) [54].

Machine learning can be divided into three different categories

- Supervised learning
- Unsupervised learning
- Reinforcement Learning

Methods like classification, prediction, and regression fall under the category of supervised learning, where the user is expected to monitor the algorithms during the learning phase and make adjustments to the settings to improve the outcome. Unsupervised learning, which focuses on techniques for rule learning, compression, and clustering, leaves the fine-tuning to the algorithms.

Reinforcement learning differs from supervised learning in that it does not need the presentation of labelled input/output pairings or the explicit correction of suboptimal behaviors. Instead, the emphasis is on striking a balance between exploitation and exploration (of current knowledge)

Providing accurate data to a machine learning model is one of the difficulties in constructing one. The data are divided into two categories: training and testing data. The model is trained using training data to make accurate predictions. At the conclusion of the development phase, test data serves as a validation tool. The model will provide an estimated percentage of how effectively it can identify a random test sampling using machine learning. Additionally, the information offered should be in line with the issue that has to be resolved. A visual representation of the development process, from data preparation to model construction, is shown in figure 2.1.

A machine learning model should be able to generalize from training data and make predictions about new data pertaining to the same issue domain. The fundamental cause of a machine learning system performing poorly is over and under-fitting. A model is said to be over-fitted when it recognizes the pattern of the training data rather than learning from it[6]. This is typically brought on by a big dataset that is too complicated for the model to fit. The training score of an over-fitted model is typically higher than the validation score, which is typically lower. On the other hand, a model that is under-fitting cannot generalize to either new data or training data. A too-small dataset is typically to blame for this. A model that is under-fitted has a declining training score despite having a reasonably high validation score.

[6] https://www.ibm.com/cloud/learn/overfitting

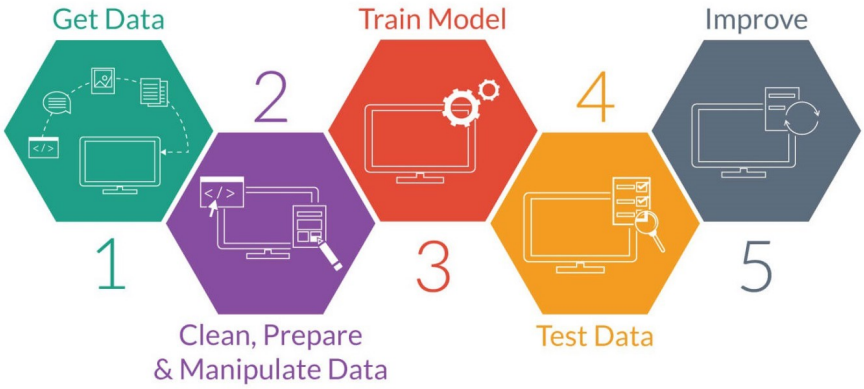


Figure 2.1. *The workflow of a machine learning process* [7]

2.3.3 Deep learning

A larger family of machine learning techniques built on artificial neural networks and representation learning includes deep learning. In fields like computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection, and board game programs, deep-learning architectures like deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks have been widely used. These applications have led to results that are comparable to and in some cases even better than those of traditional approaches, LeCun et al. (2015) [36].

Artificial neural networks (ANNs) were developed as a result of biological systems' dispersed communication and information processing nodes. Biological brains and ANNs differ in a number of ways. The biological brain of the majority of living animals is dynamic (plastic) and analog, in contrast to artificial neural networks' tendency toward static and symbolic behavior.

Deep learning refers to the employment of several network layers, as indicated by the term "deep." Deep learning focuses on an unbounded number of layers with bounded sizes, allowing for practical application and optimal implementation while maintaining theoretical universality under simple circumstances, LeCun et al. (2015) [36].

[7] https://www.muycomputerpro.com/2019/03/07/machine-learning-cuando-aprenden-las-maquinas

Chapter 3

Techniques and methods

In this chapter, we provide the most important knowledge for readers to understand this work, giving an overview of each method that was used in this thesis. Additionally, evaluation metrics and software details are discussed.

3.1 Machine Learning and Statistical Methods

3.1.1 Linear regression

Simple linear regression is a model with a single regressor x that has a relationship with a response y that is a straight line. This simple linear regression model can be expressed as:

$$y_i = a + \beta x_i + \epsilon_i$$

The linear dependency of the dependent variable Y from the independent variable X is best expressed by the parameters a and β. The difference between the actual value of Y and the predicted value of Y is what is known as the random variable which indicates the regression error ϵ, Brook and Arnold (2018) [48].

The term "multiple linear regression" refers to the case where the dependent variable Y is linearly dependent on more than one independent variables $X(X_1, X_2, X_3, ..., X_n)$. The general form of the equation for the relationship between these independent variables (X) and the dependent variable (Y) is:

$$y_i = a + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon_i$$

One can see that the equation for multiple linear regression follows the same format as the equation for basic linear regression. However, multiple linear regression, like the simple linear regression, depends on a number of assumptions regarding the variables being used, including that there is a linear relationship between the independent and dependent variables, that the variables are measured without error, that there is no autocorrelation, that there is little to no multicollinearity, and that there is homoscedasticity, Brook and Arnold (2018) [48].

3.1.2 Support vector regression

Support vector machines (SVMs) are binary linear classifiers that can also be used for multiclass classification problems and regression, like in our case. Building an ideal model that can classify incoming data points into one of two categories is one of the key applications. The fundamental idea is to create a hyperplane as the boundary between the two classes, which was first introduced by Vapnik and Lerner (1995) [10]. The wonderful thing about hyperplanes is that they are perfect for general solutions because they can easily be applied in higher dimensions as well.

The shapes in figure 3.1 below represent training data, where circles are of class A and triangles belong in class B. The ideal decision boundary to distinguish between data of two classes is depicted with the black line. The hyperplane classifies the data into various groups.

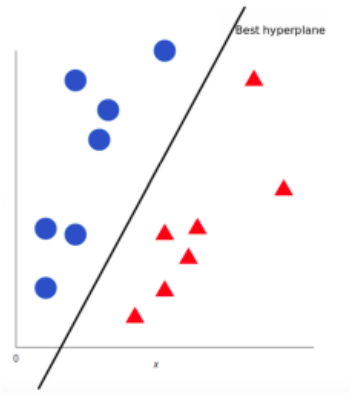


Figure 3.1. *The suggested decision boundary* [1]

This is referred as the "ideal hyperplane" in the sense that the the distance between the point nearest the hyperplane and the hyperplane itself is maximized. Even in the presence of some noise, a maximized margin will increase the likelihood that a new data point will be correctly identified.

The mathematical process of finding the optimal hyperplane can be described as:

$$\text{Minimize } \frac{1}{2}\|w\|^2 \text{subject to } y_i(w^T x_i + b) \geq 1$$

where x is the data, y the class label, $\|w\|$ the euclidean norm, w the weight vector denoting the coefficients of the hyperplane and b the bias.
Two important parts of SVMs is the kernel trick and the sparsity of the solutions. A kernel is a collection of mathematical operations that take input data and change it into the desired form. In higher dimensional space, these are typically used to find a hyperplane. Sparsity is the construction of support vectors using a subset of the training set, or more specifically using fewer variables, ignoring those with zero coefficients (no impact on the

[1] https://www.datacamp.com/tutorial/support-vector-machines-r

model). As part of the optimization objective of the SVM, $\hat{y}(x)$ can be reformulated by:

$$\hat{y}(x) = \hat{w}_0 + \sum_i a_i x_i^T x$$

where a_i is Lagrance multiplier for point x_i. The attempt to find the best hyperplane is a constrained optimization problem and is solved by the lagrangian multiplier method. The calculation of the parameters on the right part of the equation above, lets us decide the class y of any point x.

Even though this is typically the case when working with real data, not all data sets can be separated in a linear fashion. The above algorithm to maximize the margin has a clever trick in that it still functions in higher dimensions. If we encounter nonlinear separable points in the x space, we can perform a nonlinear transformation into a much higher dimensional space and use the linear SVM method to solve the problem there. Once we've found the answer, we may transform the linear hyperplane back into the x-space, where it will appear as a "snake" dividing the points. The support vectors with positive a in x space will be those in the higher dimension. In order to make SVM capable transformations into infinite dimensions without having to pay the computational cost of the transformation or the cost to calculate the inner product, Boser et al. (1992) [4] proposed applying a kernel trick to the maximum margin problem.

Similarly to SVMs, Support Vector Regression operates on the same principle. The core element of SVR is to identify the best fit line, in order to predict discrete values. The hyperplane with the maximum number of points is the best-fit line in SVR.

3.1.3 Random Forest

Leo Breiman (2001) [5], created the random forest technique. It is an ensemble learning methodology that combines bagged trees and the random subspace method. Bootstrapped aggregation is referred to as bagging. In this method, a group of various classifiers are trained using various sampled (sampling with replacement) iterations of the original dataset in this ensemble learning technique known as bagging. Decision trees are a common strategy for classification and regression, and bagging is particularly successful on high variance-low bias algorithms. Bagging is equivalent to the random subspace approach, which simply uses the dataset's features. These techniques are combined in one by random forests. Thus, the following steps represent the final algorithm:

- Create K different datasets from the original dataset D, using sampling with replacement.
- For each sample S select a number of N features.
- Grow a decision tree for each S. In these simple tree structures, leaves represent class labels or discrete values and branches represent conjunctions of features that lead to those labels or values.
- When predicting a class C or a discrete point C, aggregate the K predictions p of each tree. Assign C= Majority vote(P).

The fact that the out-of-bag error (OOB) provides a very good estimate of the n-fold cross validation (resampling method that uses different portions of the data to train a model on different iterations) error is an intriguing characteristic of random forests. The proportion of trees in the ensemble that incorrectly classified each training instance is taken into account when calculating the OOB. The importance of each characteristic can be determined using random forests, which integrate feature selection. Calculating a feature's average performance measure across different ensemble trees is one technique to achieve this. This could be the Gini index or another metric for classification or the root mean squared for regression or another suitable regression criterion. Because of this property, random forests are a particularly effective solution for problems with a lot of features.

3.1.4 Gradient boosting trees

Tree models are straightforward, understandable models. The model takes on a shape that can be seen as a tree structure, as the name would imply. Figure 3.2 displays one instance. The root node is the node at the very top of the tree. There are branches from this node to nodes below. Internal nodes or splits are the nodes in a tree that have branches extending from them. The terminal nodes or leaves are the nodes at the base of the tree. Unfortunately, tree models typically have poor predictive power. But they often have very excellent predicting powers when numerous tree models are coupled, as in bagged trees, Random Forests, or boosting algorithms.

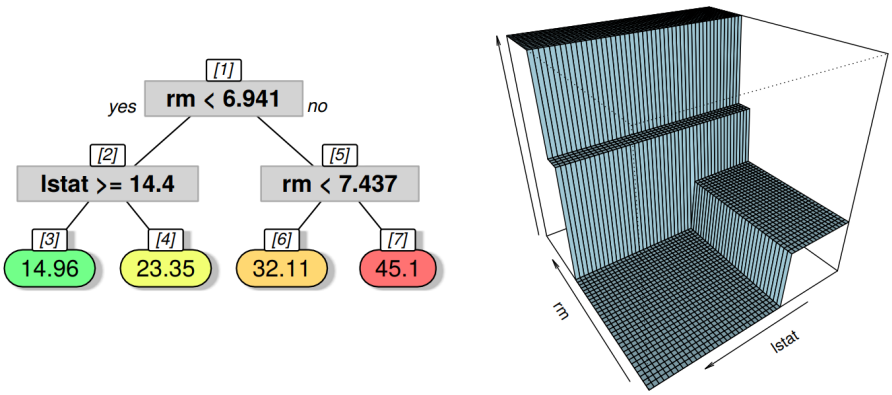


Figure 3.2. *Tree visualization and fitted function* [2]

Boosting is one of the primary methods of ensemble learning, that builds models in a sequential fashion, each one learning from the mistakes of the one before it. Each model that is trained iteratively builds on the predictions of the preceding models to create a strong overall prediction, starting with a weak base model. Gradient descent is used to find subsequent models in the direction of the average gradient of the leaf nodes of prior models, which is determined with respect to the error residuals of the loss function.

A group of boosted regression trees known as Gradient Boosting Decision Trees (GBDT) or Multiple Additive Regression Trees (MART) is an ensemble model of boosted regression

[2] https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/

trees, Friedman (2001) [16] and (2002) [17]. An ensemble of decision trees is created as the prediction model. In addition to allowing optimization of any differentiable loss function, it develops the model in stages and generalizes them. However, it has the drawback that trees added at later iterations tend to have little effect on the prediction of the majority of occurrences and just a little impact on the remainder. Due to this, the model performs poorly when applied to unobserved data and is also too sensitive to the contributions of the few, originally introduced trees.

An additive tree model, a sum of multiple trees, is shown in figure 3.3. This is fit to the same data as the tree model shown in figure 3.2. It is immediately apparent that the fit is much smoother than that of a single tree model.

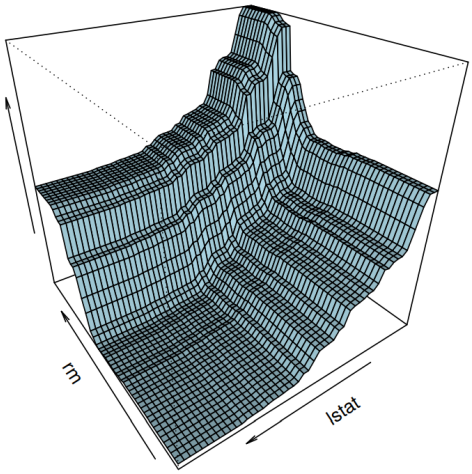


Figure 3.3. *Visualization of an additive tree model fit* [3]

Gradient-boosting creates the model in stages because it must determine the gradient after each tree is added. Thus, it should come as no surprise that the initial implementation in libraries like Scikit-Learn only makes use of one core and is not parallelized. Gradient boosting is made parallel by later algorithms like XGBoost, but not by building parallel decision trees, but rather by exploiting parallelization within a single tree Chen and Guestrin (2016) [7]. These techniques produce a parallel algorithm for split finding by collecting statistics for each column simultaneously. Additional XGBoost improvements focus on sparsity and cache awareness. Recently, XGBoost has earned a lot of fame and attention as the preferred algorithm for numerous successful machine learning teams.

Regression trees serve as the weak learners when utilizing gradient boosting for regression, and each one of them associates each input data point with a leaf that holds a continuous score. With a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity, XGBoost minimizes a regularized (L1 and L2) objective function (in other words, the regression tree functions). Adding new trees that forecast the residuals or errors of earlier trees, which are then integrated with earlier trees to produce the final prediction, is how the training process is carried out iteratively. Because the loss when introducing new models is minimized, the

[3] https://syncedreview.com/2017/10/22/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition/

technique is known as gradient boosting.

3.1.5 AutoRegressive Integrated Moving Average

AutoRegressive Integrated Moving Average is referred to as ARIMA, Lames and McGarry (2019) [34]. In statistics and econometrics, it is a well-known time-series model that is frequently employed. It is a generalization of the Autoregressive Moving Average (ARMA) model that uses differencing to address the limitation of the ARMA's suitability for stationary time-series data. Regression analysis in the form of the ARIMA model is widely used to comprehend the data and generate predictions for the future based on past results. Non-seasonal ARIMA is frequently referred to as ARIMA (p, d, q). According to Box et al. (2015) [18], the three parts are as follows:

- Autoregressive (AR): It refers to a model with a variable that regresses on its own lagged values. The number of lagged observations (i.e., the lag order) is indicated by p.

$$y_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \epsilon_t,$$

  where a_n are model parameters, c is a constant (the mean value of y in time-series) and ϵ_t is white noise. White noise means, that if we set the p parameter as zero ($AR(0)$), with no autoregressive terms, each data point is sampled from a distribution with a mean of 0 and a variance of sigma-squared and this results in a sequence of random numbers that can't be predicted.

- Integrated (I): It is the time of differencing applied to the raw data that allows the time-series data to become stationary. It is also known as the degree of differencing and is indicated by d.

- Moving Average (MA): It indicates the dependency between an observation and a residual error from a moving average model applied to lagged observations. The order of the moving average is expressed as q.

$$y_t = c + \sum_{k=1}^{q} \partial_k \epsilon_{t-k} + \epsilon_t,$$

For Seasonal Autoregressive Integrated Moving Average (SARIMA) models, they are generally denoted as SARIMA$(p, d, q)(P, D, Q, m)$, where $P, D,$ and Q are the same three terms for the seasonal part of the model, and m is the number of periods in each season.

We should also mention ARIMAX model, which we use in this dissertation and 'X' stands for exogenous variables. The ARIMA method described previously, only takes into account the target variable's past data when forecasting the present value of the target variable. There may be extra details that are very important to the forecast. ARIMAX can be regarded as a merger of the regression model and the ARIMA model. Also known as dynamic regression models, these models perform regression on ARIMA errors. The

normal regression model equation is

$$y_t = c + \beta_1 x_1 + ... + \beta_k x_k + \epsilon_t \Leftrightarrow$$
$$y_t = \sum_{j=1}^{r} \beta_j x_{j_t} + \epsilon_t$$

where y is the target output variable with n input predictor variables x_1 to x_k. The complete ARIMAX model can then be written as:

$$y_t = c + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{k=1}^{q} \partial_k \epsilon_{t-k} + \sum_{j=1}^{r} \beta_j x_{j_t} + \epsilon_t,$$

ϵ_t, as we mentioned before, is considered to be the uncorrelated white noise which can be tested using statistical tests. These ϵ_t errors in ARIMAX show autocorrelation, Hyndman and Athanasopoulos (2018) [28]. A statistical test, like Durbin-Watson, which measures the degree of lag-1 auto-correlation in the residual errors of regression can be used. A mediocre value could imply no lag-1 auto-correlation, a value closer to 0 a strong positive auto-correlation, while a bigger value implies a strong negative auto-correlation at lag-1 among the residuals errors.

3.1.6 Multi-layer perceptron

Artificial neural networks (ANNs), also referred to as neural networks, are computerized models of the connections between neurons in the human brain that are based on a network of nodes (neurons). Each neuron has the ability to take in signals from other neurons and send them to other neurons. An edge connecting two neurons has a weight assigned to it that models the significance of this neuron's input to the output of the other neuron.

However, in machine learning, the multi-layer perceptron (MLP) model is typically referred to when the term "neural networks" is used. A classification or regression model is a multi-layer perceptron. It is made up of connections between the layers' various layers of nodes (also known as "neurons"). Figure 3.4 displays an illustration of an MLP.

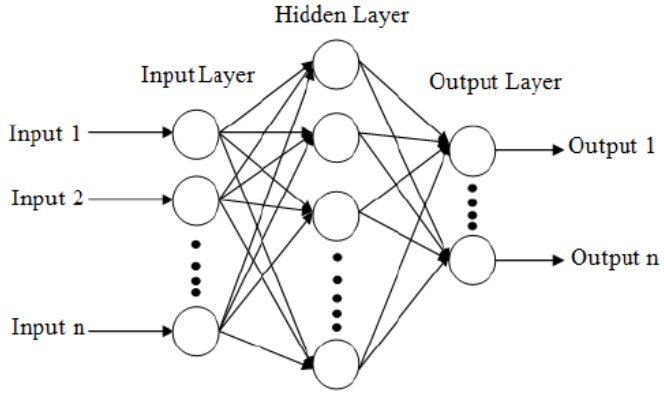


Figure 3.4. *Feedforward neural network example* [4]

It is called feedforward, because all information flows in a forward manner only and nodes' connections do not form a loop. The input layer, which is the first layer, contains the values for each feature of a dataset instance. The network's prediction is generated by the final layer, which is referred to as the output layer. One or more layers of "hidden layers" of neurons make up the middle. As can be seen below, a single neuron's activation is calculated as the sum of its inputs:

$$g = \sum_i w_i x_i$$

with g being the activation, and w_i the weight corresponding to the input x_i.
This sum is passed through a non-linear function known as the "activation function" for the hidden layers, and occasionally the output layer. Most likely, the sigmoid shown in formula below is the most typical activation function.

$$\sigma(x) = \frac{1}{(1 + e^{-x})}$$

Setting the appropriate weight between each pair of neurons in the neural network model involves training it. This is accomplished using the back-propagation algorithm, which takes in a fresh training example, calculates the gradient of the loss function with respect to the weights (a function that quantifies the error between the prediction of the neural network and the actual value), and updates the weights from the output layer all the way back to the input layer:

$$w_{ij}(t + 1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}}$$

where:

- η is the learning rate and determines the magnitude of change for the weights
- C is the cost/loss function which depends on the learning type and neuron activation functions used

The benefit of using neural networks as classification or regression models is that, in comparison to other methods, they frequently achieve a high level of predictive accuracy. However, they have some disadvantages like the need for a very large amount of training data to optimize the model.

Additionally, neural networks are not deterministic because there is no guarantee that they will converge to a single answer. Finally, Neural Networks are not interpretable since there are typically too many layers and neurons to grasp the strength of the associations between each input variable and the output variable through the various weights.

3.1.7 Long short-term memory

The fact that the input and output of a typical neural network are both fixed in length is a major restriction or limitation. The output, for example, while doing classification, is

[4] https://www.researchgate.net/figure/Structure-of-a-one-hidden-layer-MLP-Network_fig1_260321700

a probability of distinct fixed classes assuming the input categories are all the same size. Recurrent neural networks, or RNNs, attempt to solve the problem of fixed input/output sizes. RNN adds loops between actions or events, enabling the usage of information at a later time, much like a memory, Medsker and Jain (1999) [38]. A loop can be thought of as a duplicate of the network with identical parameters that just conveys the state to the following phase. This makes it possible to apply an RNN to input and output sequences that can take into consideration historical data. This has proven to be quite effective in a variety of contexts, including natural language processing, video/image classification, etc.

An RNN employs the same parameters (w, η above) throughout all steps, as opposed to a conventional deep neural network, which uses different parameters at each layer. This fact shows that the RNNs carry out the same task with various inputs at each level. The total number of parameters the RNN must learn is significantly decreased. A classic neural network can be trained similarly to an RNN, but with a little surprise added. The gradient at each output depends not only on the calculations of the current time step but also on the computations of the previous time steps because the parameters are shared across all time steps in the network.

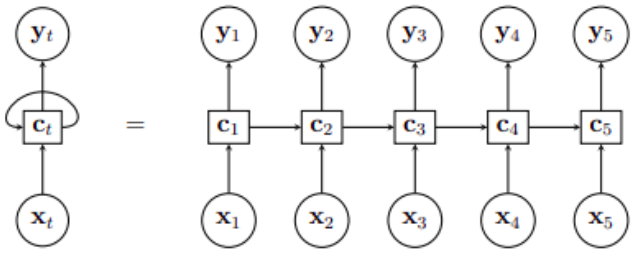


Figure 3.5. *Synced recurrent neural network, many-to-many* [5]

Figure 3.5 visualizes a synced recurrent neural network. The output at each time-step matches the input at that particular time. A loop that is a copy of the network with the same parameters simply sends the state to the following stage. This makes it possible to apply an RNN to input and output sequences that can take into account previous information.

The vanishing or exploding gradient is one major issue with a basic RNN. This may occur during training and during determining the back-propagation gradients. The chain rule, which frequently multiplies small numbers, is used to calculate the gradients, which causes the error signal to diminish exponentially. Similar effects, such as an exploding gradient, can also occur when gradients are too large. This is equally bad because the error cannot propagate effectively and the network's early layers are unable to learn, Kolen and Kremer (2001) [33].

We need to correctly initialize the weight matrices w of the RNN unit in order to solve the vanishing gradient problem's challenges. The vanishing gradient problem can be resolved with the use of appropriate regularization techniques. ReLU is a lot more

[5] https://www.semanticscholar.org/paper/Football-match-prediction-using-deep-learning-Nyquist-Pettersson/e556af01e86c3414042aa69831ea5fb398e66f94

preferable alternative to tanh or sigmoid activation functions. Since the derivative of ReLU is a constant of either zero or one, vanishing gradients are unlikely to affect it. Using either Long Short-Term Memory (LSTM) architectures is an even more well-liked approach.

Long-term input skipping and short-term information preservation have been difficult to reconcile. Hochreiter and Schmidhuber (1997) [26] presented the initial strategies to deal with this problem. A computer's logic gates served as inspiration for the design of the LSTM. Multiple gates in the design allow for management of the memory cell. We attempt to comprehend the significance of the gates and their mechanism.

A LSTM's stored value is not iteratively squashed over time because it does not use an activation function within its recurrent components. A memory c_t^j is kept in each j-th LSTM unit at time t. The result h_t^j is calculated using:

$$h_t^j = o_t^j \tanh(c_t^j)$$

where o_t^j is an output gate:

$$o_t^j = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)^j$$

where σ is a logistic sigmoid function, V_o is a diagonal matrix and W_o, U_o are weight matrices.
The update on memory cell c_t^j is:

$$c_t^j = f_t^j c_{t-1}^j + i_t^j \tilde{c_t}^j$$

and memory content is:

$$\tilde{c_t}^j = \tanh(W_c x_t + U_c h_{t-1})^j$$

and f_t^j is the forget gate, which controls how much of the memory will be forgotten. The input gate i_t^j determines how much new memory content is introduced to the memory cell. They are computed by:

$$f_t^j = \sigma(W_f x_\mathrm{l} t] + U_f h_{t-1} + V_f c_{t-1})^j,$$
$$i_t^j = \sigma(W_i x_\mathrm{l} t] + U_i h_{t-1} + V_i c_{t-1})^j$$

In figure 3.6 we see a visualization of all things presented above.

Lastly, we see the structure of a multi-layer LSTM network with 3 layers in figure 3.7. x is the input at time t and h is the output of layer j at time t. The output from layer j at time t is passed to layer $j+1$ at time t and is also fed on to the same layer j at time $t+1$.

[6] https://www.semanticscholar.org/paper/Football-match-prediction-using-deep-learning-Nyquist-Pettersson/e556af01e86c3414042aa69831ea5fb398e66f94/figure/3

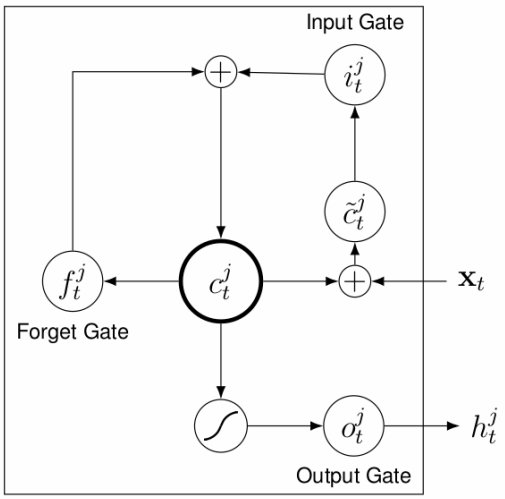


Figure 3.6. *LSTM unit* [6]

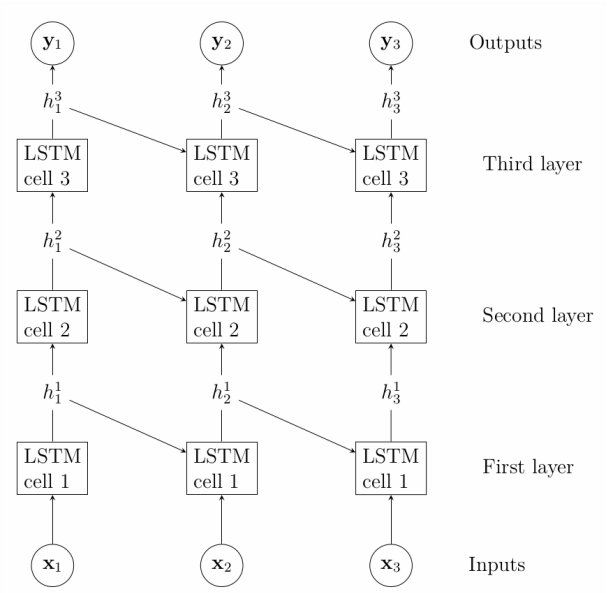


Figure 3.7. *An example of a multi-layer LSTM network with 3 layers* [7]

3.2 Evaluation Metrics

To evaluate and rate the performance of prediction algorithms, a wide range of measures have been proposed in the literature, Hyndman and Koehler (2006) [29]. There are benefits and drawbacks to interpreting the output of models that are being compared for each statistic. We chose the metrics based on their comprehensibility, applicability, and acceptance in both statistics and machine learning.

[7] https://www.semanticscholar.org/paper/Football-match-prediction-using-deep-learning-Nyquist-Pettersson/e556af01e86c3414042aa69831ea5fb398e66f94/figure/4

3.2.1 Mean absolute error (MAE)

Knowing the average difference between the actual value and the predicted value using the Mean Absolute Error enables us to understand the performance of our model. The scale used by the mean absolute error matches that of the measured data. Due to the fact that this is a scale-dependent accuracy metric, it cannot be used to compare series with varying scales. The formula that calculates this error is very simple:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value by the regression model.

3.2.2 Root mean square error (RMSE)

The most widely used and simple-to-understand measure for regression type models of all kinds is RMSE. In addition, it is directly related to the cost function that is based on least squares in the majority of the models. It is definitely a good choice as a performance metric because of all these characteristics. The difference between actual and expected values is measured as the square root of the average sum of squared errors. Since it depends on scale, outliers can affect the result. More impact on the RMSE is caused by larger errors.

The RMSE of predicted value $\hat{y}_t$ for the t time period of an output target variable y_t is computed for n different time periods as the square root of the mean of the squares of the differences:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t^2)}{n}}$$

3.2.3 Coefficient of determination (R^2)

R^2 indicates the goodness of fit for a model. The R^2 parameter for regression models describes how closely the regression line resembles the input data. A value of 1 means the model fits the data exactly. R^2 calculates the proportion of the dependent variable's variation that can be accounted for by the independent variables' variance. R^2 often has a value between 0 and 1. A model is not following the data trend and is performing worse than a horizontal line when R^2 is negative.

For instance, we would like to forecast the values of the output for a time series with n observations. Real values are $y_1, ..., y_n$ and the associated predicted values are $\hat{y}_1, ..., \hat{y}_n$. Residuals are defined as $e = \hat{y}_i - y_i$. Having the mean of real values as $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, total sum of squares $SS_{tot} = \sum_i (y_i - \overline{y})^2$, regression sum of squares as $SS_{reg} = \sum_i (\hat{y}_i - \overline{y})^2$ and sum of squares of residuals as $SS_{res} = \sum_i (y_i - \hat{y})^2$, R^2 is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where R^2 is scale independent, does not relate to the observation values and is sensitive

towards the variance in observations.

3.3 Software Aspects

There are many software libraries that offer implementations of various statistical and machine learning methods. The programming language in which the dissertation was developed is Python and some of the main packages that were used for data processing and algorithms are:

- pandas
- numpy
- scikit-learn
- statsmodels
- pmdarima
- tensorflow

These were chosen since they are well-known libraries with lots of features that are accessible via forums and blogs.

Chapter 4

Dataset

This chapter discusses the landscape of available soccer datasets and the difficulties encountered when handling soccer data, which can impede the efficient creation of statistical models. Additionally, it presents the complete steps of data collection process and provides an exploratory data analysis of the created dataset.

4.1 Current Datasets in Soccer Research

Soccer is a complicated sport, and this complexity permeates all areas of the game, including the field, practice sessions, and the manager's office. With few exceptions, there are no generally recognized standards for capturing game-related information, including individual performance data.

The fact that several parties with various priorities are in charge of recording individual player statistics in soccer presents a significant difficulty. Although, post game numbers don't differ a lot between organizations, companies and teams, the rating, or in other words performance of a player is inferred differently in most cases. Interceptions, assists, goals, distance covered, successful dribbles etc are not interpreted in the same way and everyone, from his view, concludes that a performance was bad, mediocre or great.

The initial enthusiasm of the initial idea of forecasting player performance, soon was left behind since it doesn't exist a universal approach on rate individual athletes in team sports like soccer. While there are many methods that some companies and teams use, it would be difficult to get access on such data. There are numerous databases that contain match statistics (score, competition,place, referee, odds, result) like football-data.co.uk/ and footystats.org/ or player statistics (attributes, age etc) like fbref.com/en/. There are also many manually created datasets available on Kaggle [1] [2] [3] [4], that contain match data scattered across different websites, fifa videogame rankings, injuries, and tweets related to soccer. Fantasy games like fantasy.premierleague.com/ and gaming.uefa.com/en/uclfantasy/ have a pointing system that measures individual performances, but contains some biases and relies in a few event statistics and they could not be taken into account for a problem like the one formulated here.

[1] kaggle.com/datasets/hugomathien/soccer
[2] kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset
[3] kaggle.com/datasets/eliesemmel/soccerplayersinjuries
[4] kaggle.com/datasets/eliasdabbas/european-football-soccer-clubs-tweets

Two of the most used soccer applications contain rating systems for players. In the first one, more than 300 distinct Opta stats are used to determine the FotMob player rating for each match. Any match the participant was a part of—even if he entered as a late substitute—counts as a match in the per-match ratings. To be counted in the rankings, a player had to have participated in at least half of all games and 90 minutes (for the per-90 numbers).

The second one, which we will use on this thesis to create the dataset is SofaScore. In 2015, they tried to answer to the growing need to quantify player performances, by using 1,500 different events in every fixture, translated into numbers of what happened on a soccer pitch. To collect statistics on events like possession, passes, duels, tackles, runs, interceptions, shots, and others, human operators manually record the data. In order to summarize players' performances in a way that is self-explanatory, their goal was to extract relevant information from these stats [5].

For a full year, their team of data analysts worked to examine hundreds of games, determine critical performance metrics, and assign values to each one. They looked at soccer as a dynamic, complex system in which each player plays a unique role. A sophisticated algorithm produces player ratings on a scale of 1 to 10 based on the actions of the players while they were on the field of play.
In figure 4.1 we see a graph showing a rating distribution.

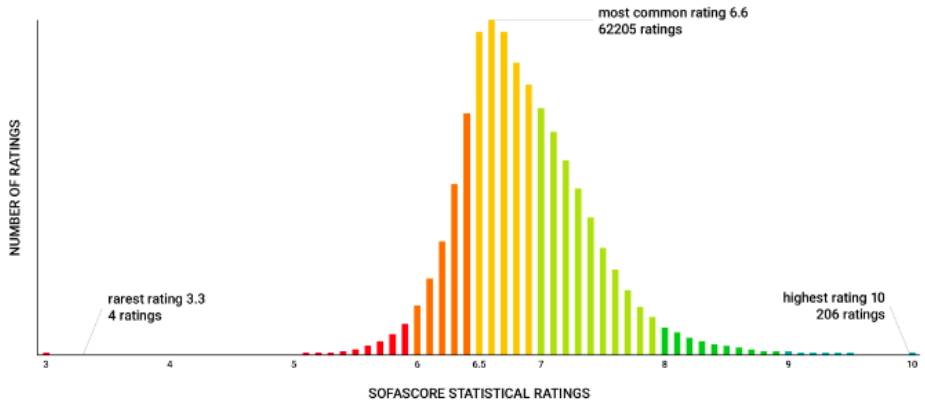


Figure 4.1. *SofaScore statistical rating distribution based on more than 690.000 ratings* [5]

As they describe in the same blog post soccer is essentially a goal-oriented sport, and if you assist in a goal, your contribution will be given greater weight in the final SofaScore ranking. But soccer is more complicated than just scoring goals, and they have created an algorithm to offer every player the same chance to receive a high rating. Therefore, in theory, a player who scores two goals but otherwise did not participate in the game and had few touches may be given a better rating than a guy who made a number of tackles, critical passes, dribbles, and interceptions and was generally all over the field.

As a machine crunches thousands of statistics to determine the grade, this is clear

[5] sofascore.com/news/sofascore-player-ratings/

sports science that leaves no room for bias. Soccer has evolved into a data-driven market, and that kind of statistical player performance ratings are quickly taking over as the industry standard.

Sticking with the above rating system and using it as the de facto way to quantify individual performances it can cause different problems. Specifically, if these calculated ratings are in some cases false, our response variable will be wrong and will result to biased conclusions about our work. However, in this thesis and after the whole research about player evaluation methods, we assume that these ratings are the best possible and will be used to study their future values.

4.2 Data Collection

After all this search for data that would may be a proper fit for answering our problem we need to collect them with various ways. Below (figure 4.2) we visualize the overview of this process

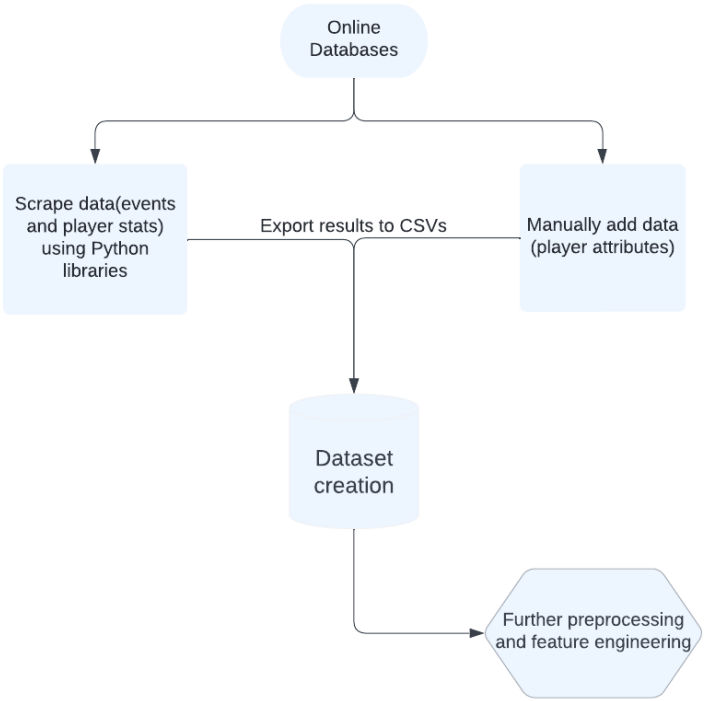


Figure 4.2. *Data collection overview*

After we selected SofaScore and other websites for getting useful data, we can get access to them in various ways, either by using web scraping or by hand, which of course seems an exhaustive process. Extraction and join of data is done in tabular format in the comma-separated value (CSV) files. The created dataset will then be available for preprocessing and preparation for algorithms. Below, we describe in detail the steps of this process.

- Firstly, we must decide what kind of players we want in the dataset. Since, the

performance metrics and match information are more established in the well known championships like Premier League or Bundesliga, we know that we will concentrate on elite players. This type of players must be further chosen in a kind of a group, either from the same club or national team or those nominated for a prize like Ballon d'Or [6]. We finally select the 30 nominees for Ballon d'Or 2021 [7] that were announced on 8 October 2021 and they are considered the best players in the world last year. The list of players in alphabetical order is:

- Nicolo Barella (ITA/Inter Milan)
- Karim Benzema (FRA/Real Madrid)
- Leonardo Bonucci (ITA/Juventus Turin)
- Giorgio Chiellini (ITA/Juventus Turin)
- Kevin De Bruyne (BEL/Manchester City)
- Ruben Dias (POR/Manchester City)
- Gianluigi Donnarumma (ITA/AC Milan)
- Bruno Fernandes (POR/Manchester United)
- Phil Foden (ANG/Manchester City)
- Erling Haaland (NOR/Borussia Dortmund)
- Jorginho (ITA/Chelsea)
- Harry Kane (ANG/Tottenham)
- N'Golo Kanté (FRA/Chelsea)
- Simon Kjaer (DAN/AC Milan)
- Robert Lewandowski (POL/Bayern Munich)
- Romelu Lukaku (BEL/Chelsea)
- Riyad Mahrez (ALG/Manchester City)
- Lautaro Martinez (ARG/Inter Milan)
- Kylian Mbappé (FRA/PSG)
- Lionel Messi (ARG/FC Barcelona and PSG)
- Luka Modric (CRO/Real Madrid)
- Gerard Moreno (ESP/Villarreal)
- Mason Mount (ANG/Chelsea)
- Neymar (BRE/PSG)
- Pedri (ESP/FC Barcelone)
- Cristiano Ronaldo (POR/Juventus and Manchester United)
- Mohamed Salah (EGY/Liverpool)

[6] francefootball.fr/ballon-d-or/palmares/
[7] lequipe.fr/La-liste-complete-des-nommes-pour-le-ballon-d-or-2021/

- Raheem Sterling (ANG/Manchester City)
- Luis Suarez (URU/Atlético de Madrid)

- Now, we want all the relevant match information, along with performance ratings of course, for all of them. We get these details from SofaScore, where the interface of the website is shown in image 4.1.

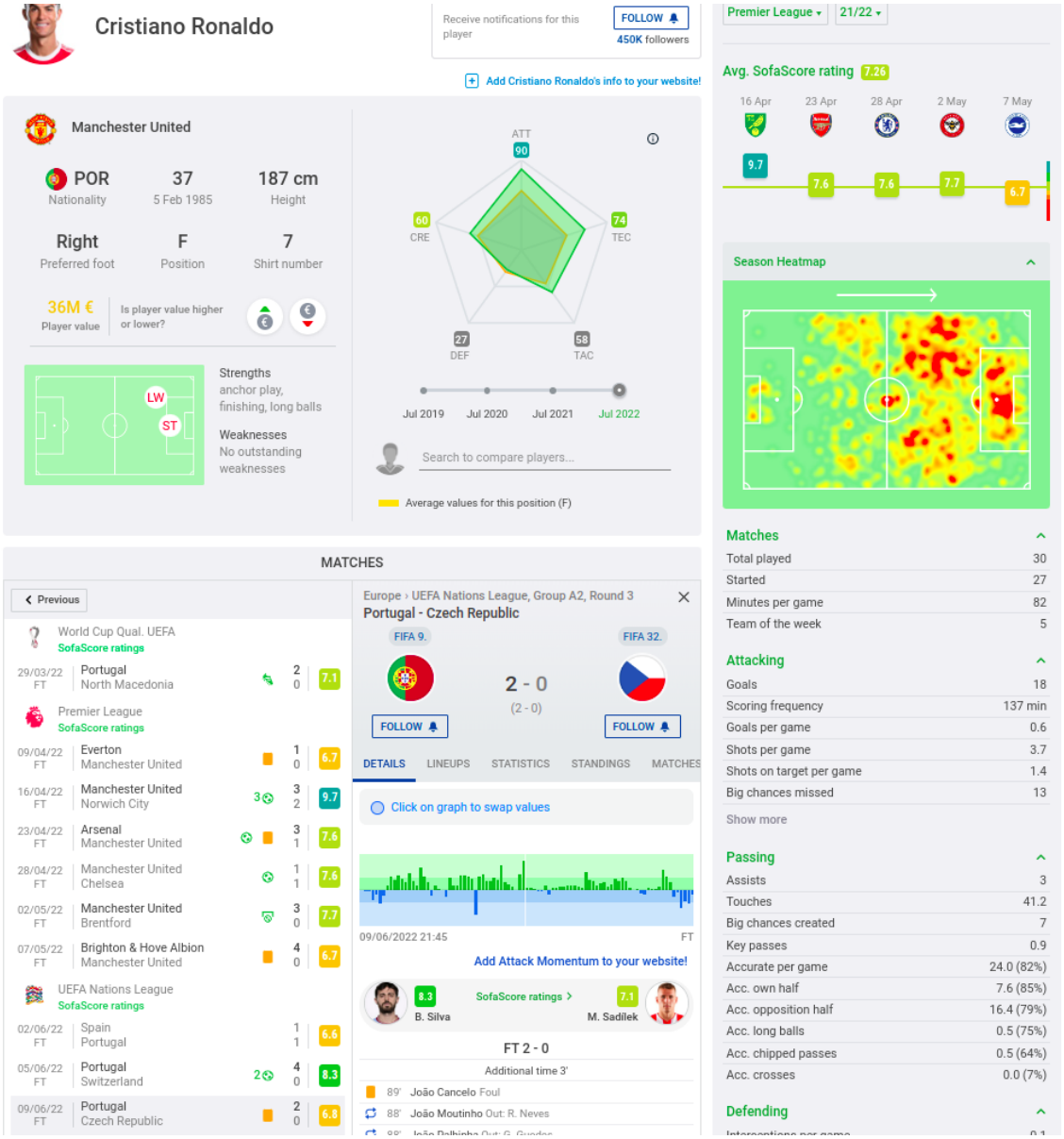


Image 4.1: *SofaScore interface* [8]

Down on the left, we see the ratings of the latest matches and it's the part that we'll focus on scraping. The format of match ratings is in json format. The same applies to event(match) information like teams, location, final score and more. So, for every one of the 30 players we retrieve all the games they played from August of 2015 until April of 2022. Notice, that before 2015 SofaScore did not assign ratings to players.

[8] https://www.sofascore.com/player/cristiano-ronaldo/750 Accessed: June 2022

- After that, we have a table where rows represent the number of total games played by every player, and columns show game information. Number of columns is 102, because along with rating, home team, away team etc, it includes useful codes and words for SofaScore API (application programming interface) and all relevant event info. Information include tournament id, tournament category, sport flag, country, round, status code(if interrupted etc), gender, teams abbreviations and colors, normal and extra time scores and injury times, number of changes, red or yellow cards and more.
  We only need some of these attributes. Specifically, we hold:
  - Performance rating
  - Matchid
  - Start timestamp
  - Tournament name
  - Round
  - Home team
  - Away team
  - Home score
  - Away score

We have collected a significant amount of information, but since soccer is such a complicated sport, we must limit the target data that the prediction methods will focus. An important consideration is the usage of games of national leagues, leaving aside international games like European Championship or World Cup. International club matches like Champions League or Copa Libertadores are also not considered in our dataset and of course, friendly games of any kind are excluded. These type of games are of special conditions (e.g. may include extra time) and not scheduled regularly like national championships. Moreover, when a player joins his national team, the teammate and the staff changes. Finally, the total number of games per player is reduced, e.g. Messi's 349 games in these 7 years include 225 league games.

Furthermore, we cannot rely only on previous performance. So we continue by creating more variables that affect the performance of a player. Before doing that,

- We begin by separating each player's current team with his opponent. This is done by checking their transfer history, since most of the players have changed teams in the 7 years of the dataset. This information can be used to check whether a match is easy or difficult, because it is well known phenomenon where players perform worst when facing an opponent of greater difficulty or the opposite, Redwood-Brown et al. (2012) [46]. Now, since we examine the best of the best, we know that most of them play (or have played) in elite clubs. That's why we separate some top clubs from the rest. Severance is done by considering most valuable clubs as indicated in transfermarkt.com/marktwertetop. The 20 top clubs in April of 2022, when last accessed this page are:

- Manchester City
- Paris Saint-Germain
- Liverpool
- Chelsea
- Bayern München
- Manchester United
- Real Madrid
- Barcelona
- Atlético Madrid
- Tottenham
- Borussia Dortmund
- Juventus
- Inter
- Leicester City
- Arsenal
- Napoli
- Milan
- RB Leipzig
- Everton
- Bayer 04 Leverkusen

The created categorical variable, with the name `'current_team_category'` or `'opponent_category'` is 1 for top clubs and 0 for all the others.

- Another dummy variable that is created by checking all matches, declares if the game is at player's team home or away. Attribute `'home_fixture'` has value equal to 1 if player plays at home stadium, and 0 of he plays at the opponent's stadium.

- Additionally, it would be useful to know the fatigue of players in every match. For that, we check date of previous games and create a new numerical variable named `'rest_days'`, which contains the number of days passed from previous match to the current one. As previous match, we do not only consider league games, where rest days are 6-7 on average, but all kind of games (international, national cups etc) where average rest varies. This feature could indicate the fatigue of athletes.

- Injury recovery also plays a major role in players' performance. That's why we manually collect injury history for all 30 players in these 7 years. The source is again TransferMarkt (transfermarkt.com/lionel-messi/verletzungen/), where injury history, suspensions and absences are based on a variety of media reports and uefa recordings. Here, three more features are created:

  - `'after_injury'`: equals to 1 if it's the first game after an injury and return was recent (less than a month), 0 otherwise.
  - `'injury_days'`: declares the number of absence due to the injury.
  - `'injury_type'`: states the type of injury, like ankle sprain, knock, flu and more.

- Another factor that affects player performance during the years is his skills like dribbling, mentality, defending, shot power, speed, stamina and so many more. These attributes are well measured and recorded in FIFA videogame, but we will only use average rating of all these and potential(the ability to become better), for every year and create two new features:

  - `'fifa_rating'`
  - `'fifa_potential'`

  The source is this dataset from Kaggle[9] and sofifa.com, and the assignment of overall rating and potential is done manually for each one of the 30 players.

- Player specific attributes that may (or may not) affect performance are also added to the dataset. These include:

  - Position
  - Foot
  - Nationality
  - Height
  - Birth

- Current age of a player in every match through the years maybe indicates differences in performance. It can be easily calculated using players' birth and games' timestamp.

- More attributes like season(e.g. 15-16, 20-21) and result(win/draw/lose) are also calculated to be used in the analysis, although post game statistics like result will not be fed to model training."

4.3 Dataset Overview and Analysis

Following the collection and the first preparation of the relevant information, the created dataset consists of 5631 games (of 30 players) and 26 features, which can be separated in four categories:

- Player attributes: name, position, nationality, birth, foot, height, current age, fifa rating, fifa potential, after injury, injury days, injury type, rest days

- Pre-Game information: match id, season, timestamp, tournament, home fixture

[9]Last accessed in April of 2022

- Team details: current team, current team category, opponent, opponent category,
- Post-Game information: home score, away score, result, performance

Below, in table 4.1, we provide a short description of each attribute, because we'll refer to them later several times.

Table 4.1. *Dataset description*

Variable	Description
player name	The name of each player for this match
player position	Position of each player
player nationality	Nationality of each player
player birth	Birth date of each player
player foot	Player's strong foot (Right or Left)
player height	The height of each player
age	Age of the player on match date
fifa rating	Overall fifa ranking for this season
fifa potential	Potential growth of ranking through this season
after injury	Whether a player came back from injury or not
injury days	Days of the injury
injury type	Type of the injury
rest days	Days passed since last game
season	Season in which the game takes place(August-May)
matchid	Unique id of the match
startTimestamp	Time and date of the match
tournament name	Name of the national league
current team	Current club where the footballer plays
current team category	Market value of the current club
opponent	Opponent side that the player faces
opponent category	Market value of the opponent club
home fixture	Whether the game is played in home stadium or not
homeScore	Number of goals of the home team
awayScore	Number of goals of the away team
result	Win, draw or lose (for player's team)
Performance	How the athlete performed on this match, based on rating

Although the fourth category of post game information will not be used in modeling to predict performance, we can explore all attributes to have a better view of the data.

Firstly, in the 30 players there are 16 Forwards, 8 Midfielders, 5 Defenders, and 1 Goalkeeper, showing that goals, assists and generally attacking is considered more important than defence and gets more traction.

Concerning players' statistics we visualize in figure 4.3 the total games played in these 7 years.

Mohamed Salah has played the most games, with Romelu Lukaku and Cesar Azpilicueta following. Pedri and Foden may be in last positions, although they are very young and

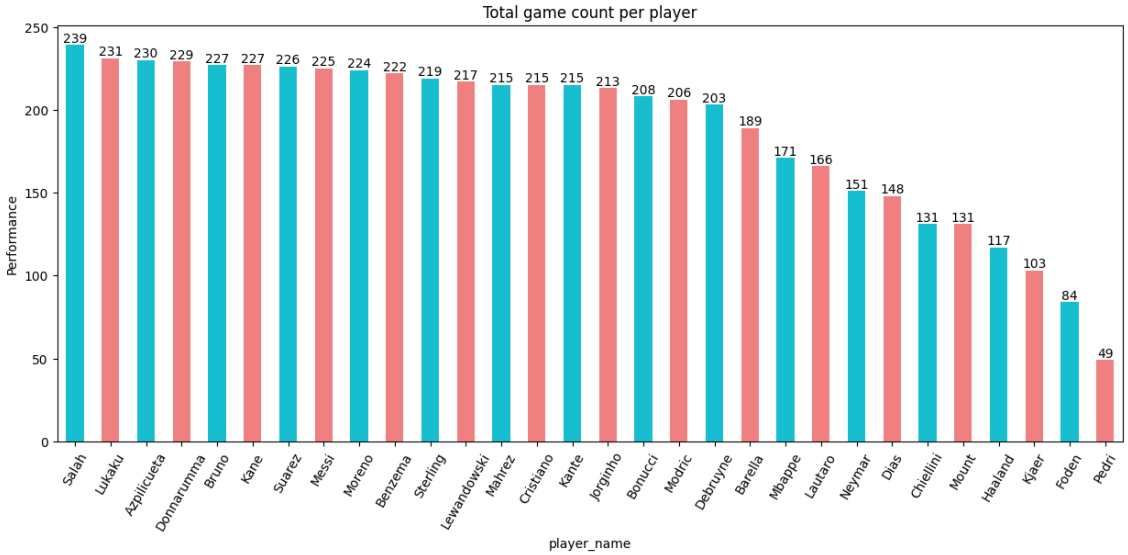


Figure 4.3. *Number of league games per player (2015-2022)*

have 2 and 3 years of professional football respectively. Neymar, is a player that worths to be mentioned. While, being a top class player all these years, he suffered from many injuries, making him lose several league games and that's why his total games are 151.

Concerning performance now, we plot the average performance for each player through this period in figure 4.4.

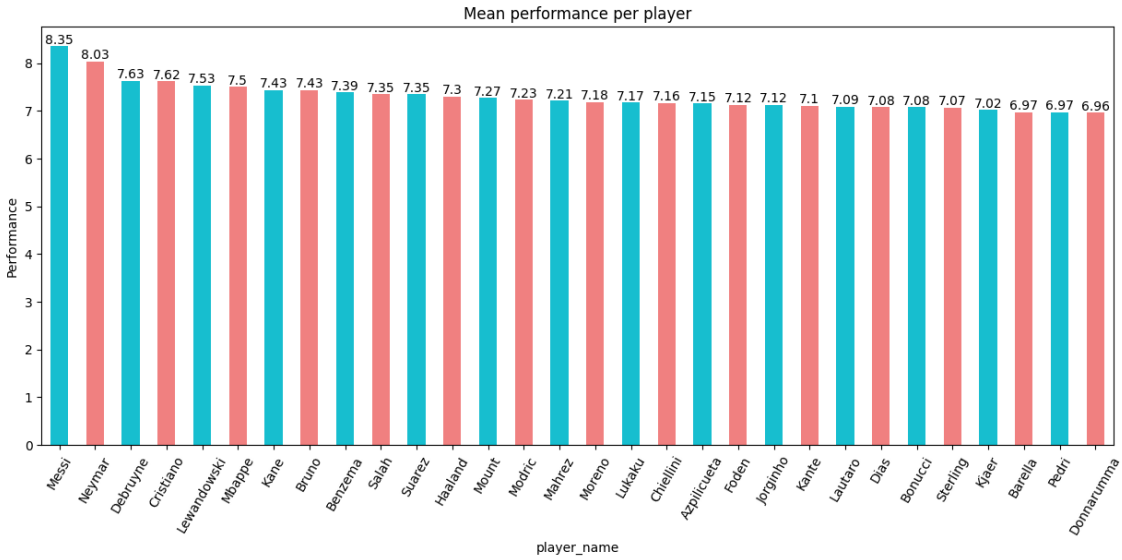


Figure 4.4. *Average performance per player (2015-2022)*

As we expected, Lionel Messi with the most Ballon d'Or wins in this period has the greatest value (8.35) and along with Neymar (8.03) they are the only players that overpass 8. De Bruyne, Cristiano and Lewandowski fill the first 5 places. Kane, in the seventh place it could be described as a surprise, since his high average performance shows stability through the years, which has not been rewarded by the media or by making a transfer to bigger club. On the other side, Donnarumma is on the last place, while being the best

goalkeeper in the world according to Ballon d'or, which indicates a possible problem and bias on current rating systems. The distribution of ratings is around 7.31 with a standard deviation of 0.81, see figure 4.5. One could see the peak of almost 500 ratings with value equal to 7.5, which may happen due to a round calculation when certain events happen.

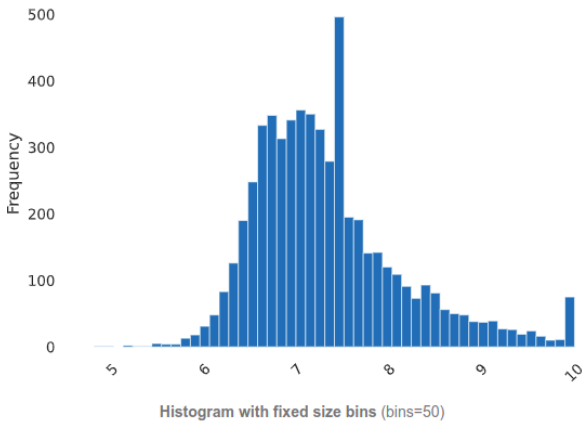


Figure 4.5. *Distribution of player ratings*

As we mentioned earlier, attacking contributions are more significant than defending ones on current performance evaluation systems. In figure 4.6, we visualize the distribution of the four soccer positions (Forward, Midfielder, Defender, Goalkeeper).

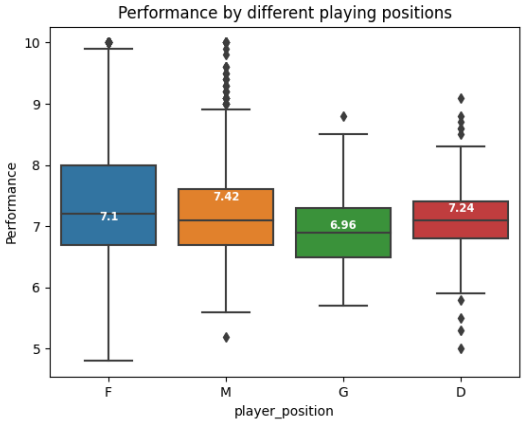


Figure 4.6. *Performance by different playing positions*

The skewness is much bigger in attacking players and the mean is lower than midfielders or defenders. The reason is that while the best performers are forwards, there are many players in this list that had lower ratings the previous years making the group average decrease.

We can see three examples of performance (blue lines) through the years for three playing positions, along with the rolling averages (red lines) of last 10 games. Figure 4.7 plots the performance of a defender, figure 4.8, shows all performances of a midfielder and figure 4.9 visualizes the performance of a forward.

On defender's (Azpilicueta) historical performances the variance is small (0.24), showing the stability of players in this position, while on midfielder(Modric) the variance seems

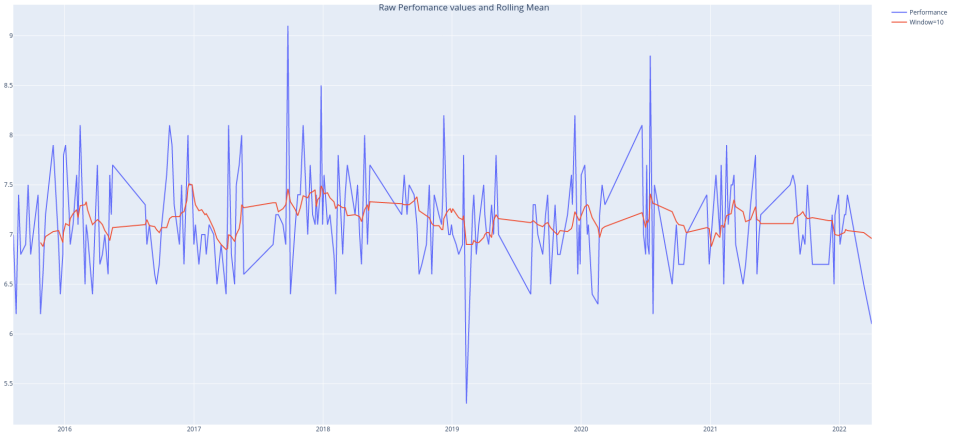


Figure 4.7. *Performance history of Cezar Azpilicueta (defender)*

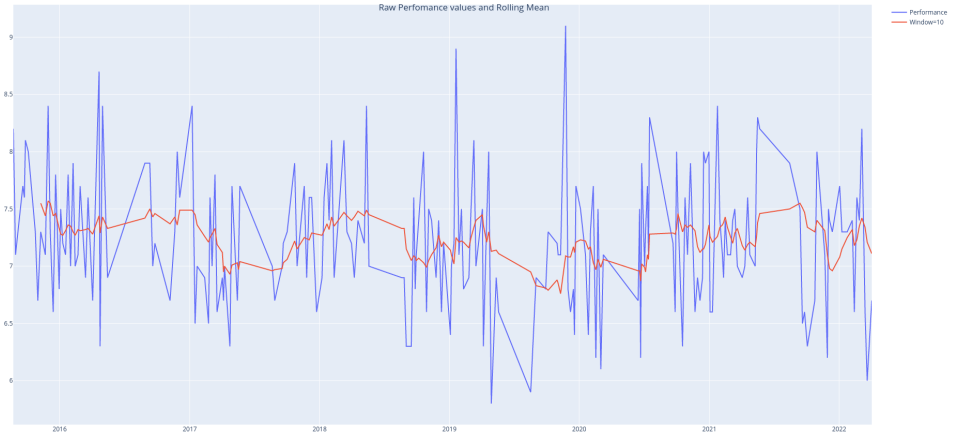


Figure 4.8. *Performance history of Luka Modric (midfielder)*

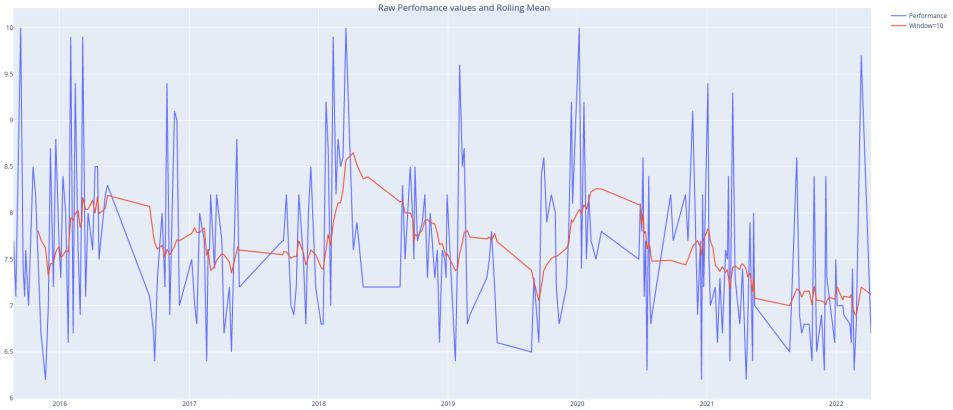
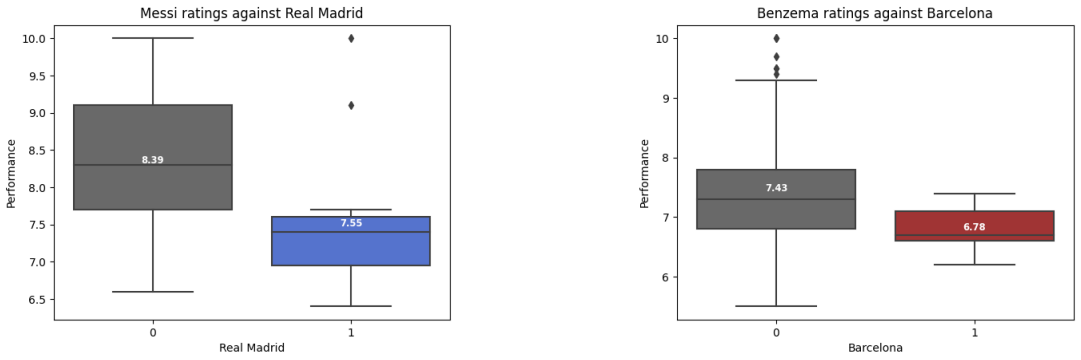


Figure 4.9. *Performance history of Cristiano Ronaldo (forward)*

a little bigger (0.35) and on forward(Cristiano) the instability in ratings is greater(*variance* = 0.74), due to the way goals and assists affect rating in positive or negative way each time. Moreover, these conclusions are easily extracted by looking the red lines of moving average for each player, where the defender has a smoother line.

Additionally, the rating time series of footballers are not periodic with respect to time, making them irregular spaced. It is clear that there are significant gaps between the three players. Injuries, summer vacations, and of course the pandemic quarantine in 2020 are to blame for the longer irregular gaps between matches. In this study, we adopt the perspective that all games occur at predictable intervals of time, as we'll see in next chapters.

Another interesting exploration would be to check how some players perform against specific teams or when the play with certain teammates. For the first case, a good rivalry is the Spanish El Classico and indicative examples are Benzema and Messi who almost have not missed a game, if we exclude Messi's last year departure for PSG. Plots in figure 4.10 show the decrease in performance in more than 10 games played on these 7 years. Both players have a reduced average, 8.39 to 7.55 and 7.43 to 6.78 respectively, when playing with their big opponent.

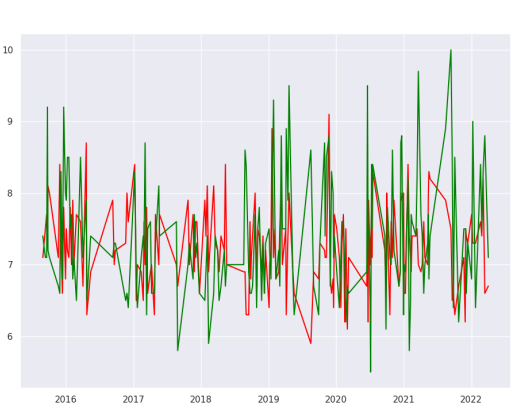


(a) *Messi against Real Madrid(1) or not(0)* **(b)** *Benzema against Barcelona(1) or not(0)*

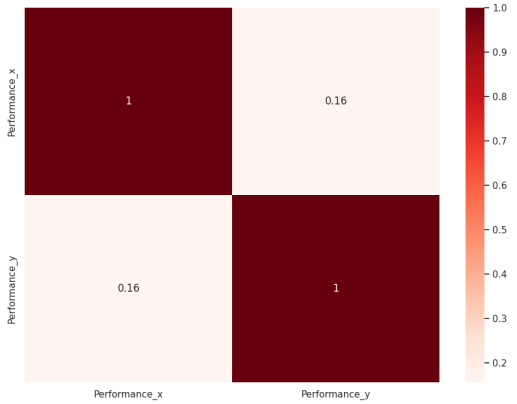
Figure 4.10. *Elite player competence against big rivals*

For the second case, a good duet to examine is Benzema and Modric, who are the only teammates appearing on this list and have not changed a club the period we examine on our dataset. Figure 4.11 shows the time series of performance in their common games and the correlation of both.

While, from the first plot we cannot infer a strong relationship, the small positive correlation number on the heatmap on the right indicates that the absence of Benzema affects Modric and the team in general, which is also confirmed by the fact that the Croatian player has an average of 7.25 in 181 games played with the French, which drops to 7.00 in 25 games played without him.



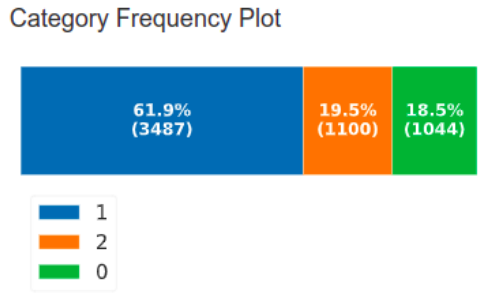
(a) *Modric(green) and Benzema(red) common games*



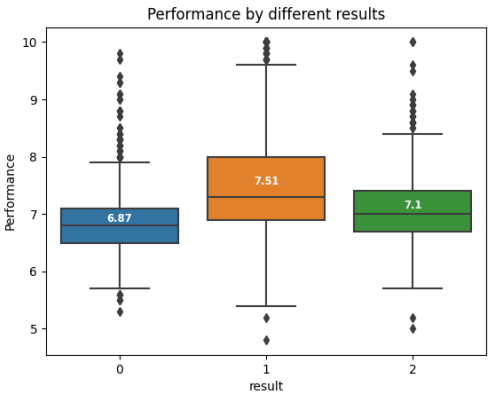
(b) *Heatmap of Modric's performance(x) with Benzema(y)*

Figure 4.11. *Modric and Benzema performance correlation*

Next, we can see how rating is affected from outcomes. Figure 4.12a shows the frequency distribution of the three results, while 4.12b visualizes boxplots of performance compating to win(1), loss(0) or draw(2).



(a) *Distribution of win(1), loss(0) and draw(2)*

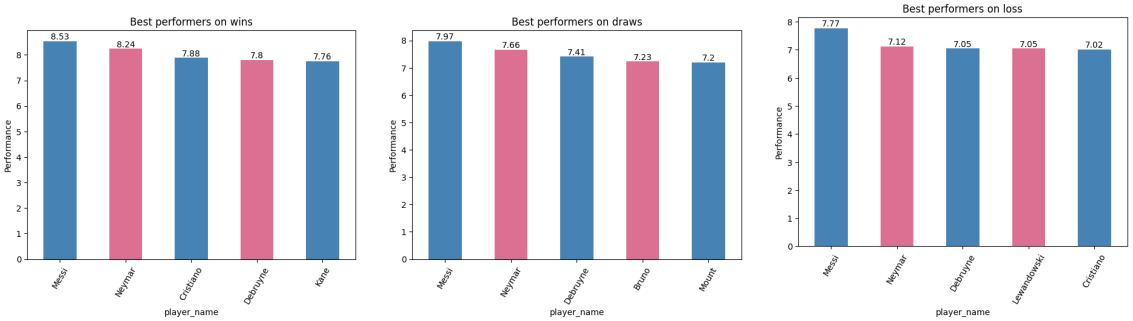


(b) *Performance distribution on different results*

Figure 4.12. *Results and performance comparison*

As it was expected, ratings are bigger on successful results. When a team performs well, the individual ratings will be bigger, reflecting the contributions of each one to the win. These contributions are goals, assists, clean sheets, interceptions and more. Let's visualize now the players that perform best on each of the three cases, figure 4.13.

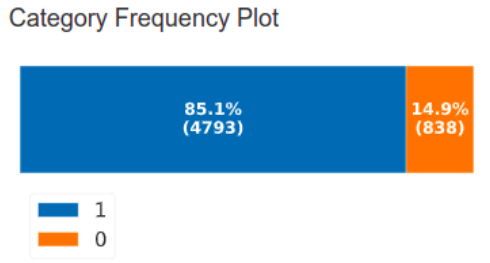
The presence os Harry Kane on wins top-five is worth mentioning, along with Lewandowski presence on loss top-five, showing that the second may score or play well even on unsuccesful outcomes.



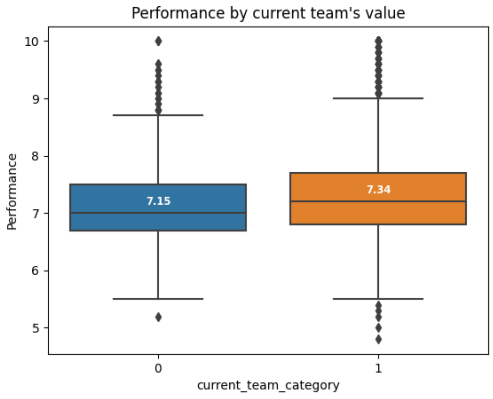
(a) *Top five players on wins* **(b)** *Top five players on draws* **(c)** *Top five players on losses*

Figure 4.13. *Best players on different outcomes*

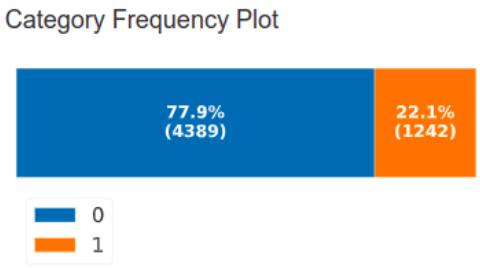
Continuing the analysis of the created variables, we plot the distributions of club categories and the correlation with the performance of the athletes, figure 4.14



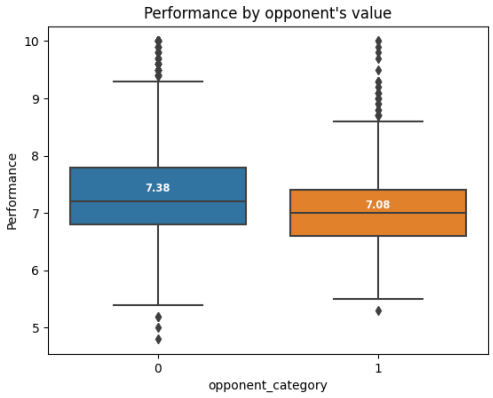
(a) *Distribution of player's team category: Elite(1) or not(0)*



(b) *Performance distribution on current team category*



(c) *Distribution of opponent's category: Elite(1) or not(0)*



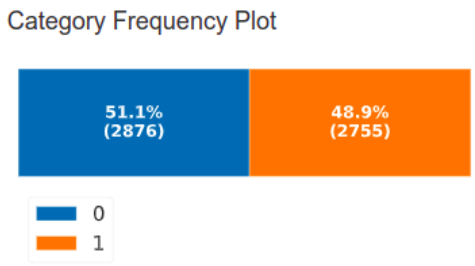
(d) *Performance distribution on oppontent's category*

Figure 4.14. *Player teams and opponents club values compares to performance*

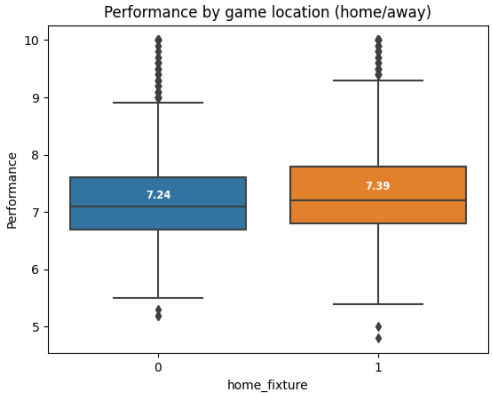
It is easily noticed that most players of our dataset play, or have played, in elite clubs, facing mainly smaller-value clubs. Additionally, since soccer is a demanding teamwork sport, it couldn't be different than the rating to be better when an athlete is part of a great

team and faces a weak side. This thing is indicated from the boxplots on the right.

Another factor that can affect player's competence is the location of every game. Specifically if the game takes place in home stadium or away. We visualize these relationships in figure 4.15



(a) *Distribution of match location, home(1) or away(0)*



(b) *Performance distribution on match location*

Figure 4.15. *Match location and performance comparison*

While this attribute is almost perfectly balanced, we observe that performance tends to be a little bigger in home events. In numerous athletic situations, the impact of game site on performance has been studied. Soccer game locations have a favorable impact on secondary and tertiary level performance, Diana et al. (2017) [12]. Game strategies change substantially from match to match, with home matchups being more structured and varied than away encounters.

Moreover, the number of rest days between games is worth to be mentioned, figure 4.16.

Value	Count	Frequency (%)
3	1194	21.2%
4	1069	19.0%
7	884	15.7%
6	571	10.1%
5	530	9.4%

Figure 4.16. *Rest days between games*

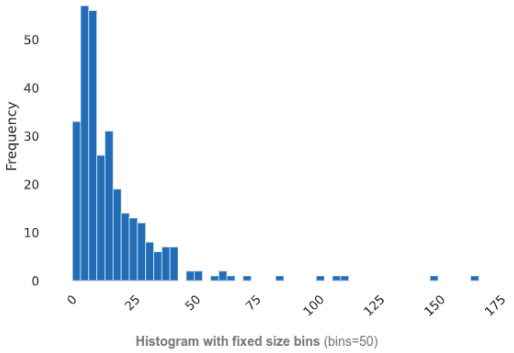
It's not an unusual phenomenon that elite players play two matches per week, including International competitions and national cups. That's why most rest durations are for 3 or 4 days, altough there exists the regularity of league games in the third place of frequencies.

Last but not least, is important to take an idea of injuries landscape on elite players of the dataset. Of the 8443 total games(europe, cups, league included) since August 2015, 304 of them were played after an injury recovery. Figure 4.17 visualizes basic information collected and adapted to the dataset.

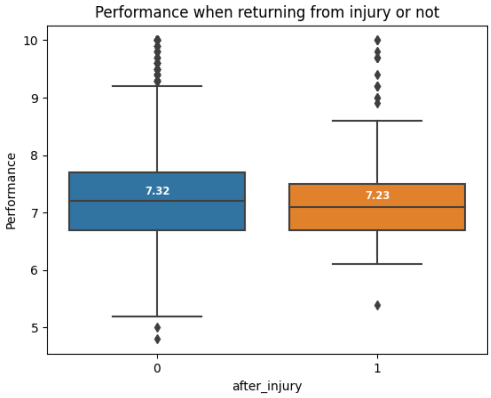
Common Values

Value	Count	Frequency (%)
Hamstring Injury	33	10.9%
Muscular problems	19	6.2%
Muscle Injury	18	5.9%
Adductor problems	18	5.9%
Ankle Injury	17	5.6%
Corona virus	17	5.6%
Knee Problems	14	4.6%
Thigh Problems	9	3.0%
Knock	8	2.6%
Calf Injury	8	2.6%
Other values (73)	142	46.7%

(a) *Most common injuries*



(b) *Distribution of recovery days*



(c) *Performance evaluation after injury(1) recovery or not(0)*

Figure 4.17. *Graphical information of injuries*

Hamstring injury, a strain or tear to the tendons or large muscles at the back of the thigh, is the most common one, with muscular and adductor problems following. Notice the presence of 17 corona virus cases, due to the recent pandemic (figure 4.17a). Average absence days due to injury are 17 with a maximum of 168 (figure 4.17b). This number is affected by the severeness of the injury, the age of the players, club's rehabilitation procedures and more. Finally, figure 4.17c indicates that performance is not greatly affected in the return of a player, although the ratings are more conservative and the outstanding performances (e.g. near 10) happen on their regular forms.

Chapter 5

Player Performance Prediction with Regression

This chapter tries to answer some parts of the problem defined in the beginning. The main goal of the following experiments is to predict the performance of a player in a random future match. Additionally, it interprets which match specific conditions are important and affect the performance of the athlete.

5.1 Dataset and Preprocessing

In the previous chapter we saw that the whole dataset consists of 5631 games (samples) and four categories of attributes: player attributes, pre-game information, team details and post-game information. In this approach, the fourth category will not be used of course, since we want to predict the individual performance before the referee's whistle.

Many of the features described in table 4.1 were used, along with a new category of attributes, that contains last games' fit for every player. The process to create these values is very simple:

- For every player in the data, calculate the rolling mean of last N games.
- Assign that value to a new column.
- The first N games have null values since there weren't several previous matches to calculate the moving average.
- Fill these empty values with the average of all performances of the player.

Moreover, player position (F, M, D, G) is encoded to 4 new dummy binary variables with values of 0 and 1. Same thing is done for player foot(R, L).

Finally, it's not correct to assume all games equal for all players and fit a model on the input data, since each player has, for example, different average performances. Every player could have different effect on the games of our dataset. All variables that don't vary over time have impacts that are captured by fixed effects. In other words, these fixed effect factors in the model would include everything else that does not change over time at the game level, such as player name.

That's why we create 30 new categorical variables, one for every footballer of our data. The feature equals to 1 when represents the respective player and 0 otherwise. So, the new dataset contains 5631 games, 50 input features and 1 output feature (Performance).

So, we can now present the new dataset in table 5.1. Needs to be mentioned that from the 30 new columns of player names, we only present the first and the last one, skipping the intermediate, for the sake of simplicity.

Table 5.1. *Dataset for Regression*

Variable	Description
player height	The height of each player (Numerical)
age	Age of the player on match date (Numerical)
FIFA rating	Overall FIFA ranking for this season (Numerical)
FIFA potential	Potential growth of ranking through this season (Numerical)
after injury	Whether a player came back from injury or not (Categorical)
injury days	Days of the injury (Numerical)
rest days	Days passed since last game (Numerical)
form last 1	Average player's performance on previous 1 match (Numerical)
form last 3	Average player's performance on previous 3 matches (Numerical)
form last 5	Average player's performance on previous 5 matches (Numerical)
form last 10	Average player's performance on previous 10 matches (Numerical)
current team category	Market value of the current club (Categorical)
opponent category	Market value of the opponent club (Categorical)
home fixture	Whether the game is played in home stadium or not (Categorical)
player position F	Forward (Categorical)
player position M	Midfielder (Categorical)
player position D	Defender (Categorical)
player position G	Goalkeeper (Categorical)
player foot R	Player strong foot (Right) (Categorical)
player foot L	Player strong foot (Left) (Categorical)
player name Messi	If the game is about Messi(1) or not(0) (Categorical)
player name ...	If the game is about ...(1) or not(0) (Categorical)
player name Azpilicueta	If the game is about Azpilicueta(1) or not(0) (Categorical)
Performance	How the athlete performed on this match, based on rating (Numerical)

5.2 Feature Significance

Here, we explore the stochastic formulation of the problem. Specifically, we want to explain which attributes affect player's performance and generalize the results of this small sample of 30 footballers to the whole population of professional players.

A first look on how collected features correlate to performance was taken on the exploratory data analysis of the previous chapter. Boxplots and distributions there, indicated the effect of some categorical variables to performance. Concerning numerical

features, a method of calculating correlation could do the work. Pearson correlation coefficient for example, is the ratio between the covariance of two variables and the product of their standard deviations, and takes values from -1 to 1. Figure 5.1 visualizes the heatmap of all coefficients between numerical variables.

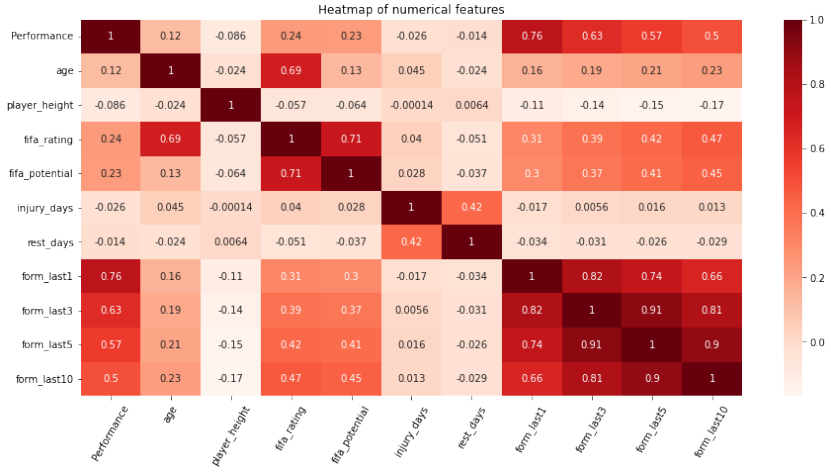


Figure 5.1. *Pearson correlation coefficients*

It is clear that the future performance of a footballer is highly dependent on previous games -correlations bigger than 0.5- proving what is mentioned as current form of a player. FIFA average skills and potentials also show a little effect with coefficients above 0.20.

However, the most straight way to check the effect of an independent variable to a dependent, is to do a statistical test and check the p-value. A p-value, also known as a probability value, is a numerical indicator of how likely it is that the data are the result of the null hypothesis. The p-value, which ranges from 0 to 1, is frequently used to indicate the statistical significance level. The smaller the p-value, the stronger the evidence that we should reject the null hypothesis.

In order to get the most important of all variables we perform backward stepwise selection (or backward elimination) method. This technique starts with a model that includes all the variables being considered (referred to as the full model), and then begins removing the least significant variables one at a time until the stopping rule is reached or there are no variables left in the model. The least significant variable is one that has the highest p-value in the model, or whose removal from the model results in the smallest reduction in AIC(or R2), or the smallest boost in Residuals Sum of Squares when compared to other predictors. When every remaining variable in the model has a p-value lower than a predetermined threshold, the halting criterion is satisfied.

In our case, a multiple linear regression model is fitted on the data. It begins with all the 50 independent variables, having performance as a dependent one and iteratively removes non significant variables with a threshold of 0.05. Fixed effect variables have big p-values, more than 0.4, and the results of the full model are not presented for space saving. The reduced model, according to this statistical test has the following features (table 5.2):

Table 5.2. *Most significant features*

Variable	coef	P-value
const	0.0397	0.635
injury days	-0.0033	0.019
rest days	0.0013	0.032
form last 1	0.9891	< 0.001
opponent category	-0.1249	< 0.001
home fixture	0.1204	< 0.001

It is indicated that the created attributes in our dissertation are important in different ways. Of course, the dynamic of the opponent plays a major role in today's soccer results. A slightly negative coefficient means that if opponent belongs to the elite clubs, the individual performance of the player would be slightly bad. Home fixture's increase by 1 unit, moderately relates with an increase in performance by 0.1204.

Moreover, the thought to consider the fatigue of a player between matches and the recovery from an injury, where he may not perform in his full capabilities, seems to be a good starting point. Increase of rest days by 1 unit, are related with a very small increase in performance by 0.0013, while an increase in the recovery duration from an injury has a small negative effect in footballer's performance.

Finally, the performance in previous games seems to be so important that an increase in last game's rating by 1 unit relates with an average increase in next game performance by 0.98.

5.3 Regression Improvement

5.3.1 First attempt

After running the statistical test with p-values from a linear regression model on the data, we can see the general performance of the model. Full model (50 variables) has a coefficient of determination:

$$\text{full model: } r^2 = 0.5924$$

revealing that approximately 60% of the variability observed in the target variable is explained by the regression model. That means that the model has quite good explanatory capabilities.

Reduced model (5 variables, table 5.2) has almost the same:

$$\text{reduced model: } r^2 = 0.5917$$

showing that the removal of the other 45 not important features doesn't affect the predictive accuracy. Simpler models are usually preferred over complex ones for three reasons [32]:

- Overfitting: can occur when a high-dimensional dataset has an excessive number of features (model captures both real and random effects).
- Interpretability: When features are connected with one another, an overly complex model with too many features may be difficult to understand.
- Computational Efficiency: A model that has been trained on a smaller dataset has better computing efficiency (execution of algorithm requires less computational time).

5.3.2 Experimentation and optimization

As a final step we try to improve regression results by doing more data preprocessing and experiment with other models, such as Support Vector Machines, Random Forest, XGBoost and Multilayer perceptron. A good way to get better results is to use data normalization, where the range of input data is scaled. This method is often required by many learning algorithms that benefit from standardization of the data set. Here, we standardize features by removing the mean and scaling to unit variance, and the standard score of a sample x is calculated as:

$$z_i = \frac{x_i - \mu_x}{s_x}$$

where μ_x is the mean of the training samples, and s_x is the standard deviation.

Moreover, we split data to train and test sets, giving a proportion of 10% to test. Empirical studies show that the best results are obtained if we use 20% of the data for testing, and the remaining 80% of the data for training, Gholamy et al. (2018) [19]. Here, where the dataset is relatively small, we choose a smaller proportion for test. After that, the shape of the two inputs are:

- training set: 5067 samples, 50 features
- test set: 564 samples, 50 features

Concerning model parameters, there was applied tuning on the hyperparameters through exhaustive search of the best ones. Random Forest, XGBoost and MLP neural network have more complex attributes to take into account and optimize. The space of parameters that used for the models in the experiments are listed in table 5.3. Detailed information about the parameters of each method can be found in the well written scikit-learn documentation.

Table 5.3. *Parameter grid for the regression models*

Model	Parameters
Linear regression	-
SVM regression	kernel=['linear', 'poly', 'rbf', 'sigmoid'], C=[0.5, 1], epsilon=[7e-2, 1e-1, 1.0]
Random Forest regression	max features=['auto', 'sqrt', 'log2'], min samples leaf=[1, 8, 15], min samples split=[2, 8], n estimators=[80, 100, 120]
XGBoost regression	learning rate=[0.3, 0.09], max depth=[5, 6, 7], min child weight=[1,5], subsample=[0.5, 1], colsample bytree=[0.5, 1]
MLP regression	hidden layer size=[(50,50,50), (50,100), (100,50), (100,)], activation=['tanh', 'relu'], solver=['sgd', 'adam'], learning rate=['constant','adaptive'], alpha=[0.0001, 0.09]

Table 5.4 presents the evaluation of the four models. R-squared, root mean squared error and mean absolute error are all measured on the unseen for the models test set.

Table 5.4. *Comparison of regression models*

Metrics	Linear reg	SVM reg	RF reg	XGBoost reg	MLP reg
R-squared	0.61297	0.60277	0.58955	0.58301	0.58013
RMSE	0.48908	0.49920	0.57520	0.51147	0.51324
MAE	0.37357	0.38319	0.43531	0.39259	0.39642

A little improvement in linear regression metrics is achieved (r^2 = 0.61), which proves the need for normalizing data to the same scale. Generally, all models have very small errors, although they perform slightly worse than the first simpler one. This fact, make us think that there are many problems that can be considered and solved with simple statistical methods. Nowadays, it is a common phenomenon to jump to very complex models of machine and deep learning, that seem more fascinating, but that is not always the proper starting point.

Chapter 6

Time Series Forecasting of Player Performance

This chapter contains approaches, experiments and results in the attempt to answer the main point of this research. This point is to forecast the performance of individual soccer players in one or more future matches. Various statistical, machine learning and deep learning methods are applied in each separate player's time series of historical performances, along with more exogenous features. Lastly and most importantly, the results of these models are compared with random predictions and naive models, in order to validate their contribution.

6.1 Dataset and Preprocessing

In the previous chapters we saw that the dataset consists of 5631 games from 30 players, nominated for Ballon d'Or award. The difference here, is that the information is broken per player to 30 smaller datasets that contain approximately 200 played games from 2015 until 2022. The number of samples (games) differs between the footballers, as is was visualized on figure 4.3

Additionally, these separate multivariate time series are not regular, concerning time. Figures 4.7, 4.8 and 4.9 show time in x axis and performance values on y axis. It is easily observed, that there are big intervals for the three players. Summer vacations, injuries and even the quarantine due to pandemic on 2020 are the reasons of bigger irregular gaps between matches. In this dissertation, we consider that all games take place in regularly spaced intervals of time.

There is no point on using some attributes of players here, since they are stable and will not offer something to the models. These attributes are name, height, foot and position. Other variables like current team's category are kept since they change for many players through the 7 years. Figure 6.1 presents a basic overview of the data used in this chapter for English super star Raheem Sterling.

This sample indicates that Sterling played 219 league games these 7 years and shows the evolution of different attributes through the years. We can observe the nine independent variables that contain numerical and categorical information, and the 10th dependent variable which is a numerical real value.

Now, a good way to check the dependence of current variable with its previous values is the autocorrelation plot, which is a statistical representation used to analyze the degree

	age	fifa_rating	fifa_potential	after_injury	injury_days	rest_days	current_team_category	opponent_category	home_fixture	Performance
2836	20.68	82	88	0	0	86.0	1	0	0	6.8
2837	20.70	82	88	0	0	6.0	1	1	1	6.7
2838	20.72	82	88	0	0	7.0	1	1	0	6.3
2839	20.74	82	88	0	0	6.0	1	0	1	7.2
2840	20.79	82	88	0	0	4.0	1	0	1	6.2
...	...	...	...	...	...	...	...	...	...	...
3050	27.20	88	89	0	0	3.0	1	0	0	9.3
3051	27.22	88	89	0	0	4.0	1	1	1	6.6
3052	27.24	88	89	0	0	7.0	1	1	0	6.5
3053	27.33	88	89	0	0	13.0	1	0	0	7.8
3054	27.36	88	89	0	0	5.0	1	1	1	6.2

219 rows × 10 columns

Figure 6.1. *Sample data of Raheem Sterling*

of similarity between a time series and a lagged version of itself. Figure 6.2, visualizes the raw time series, along with the autocorrelation plots and partial autocorrelation plot of the midfielder of Manchester City, De Bruyne.

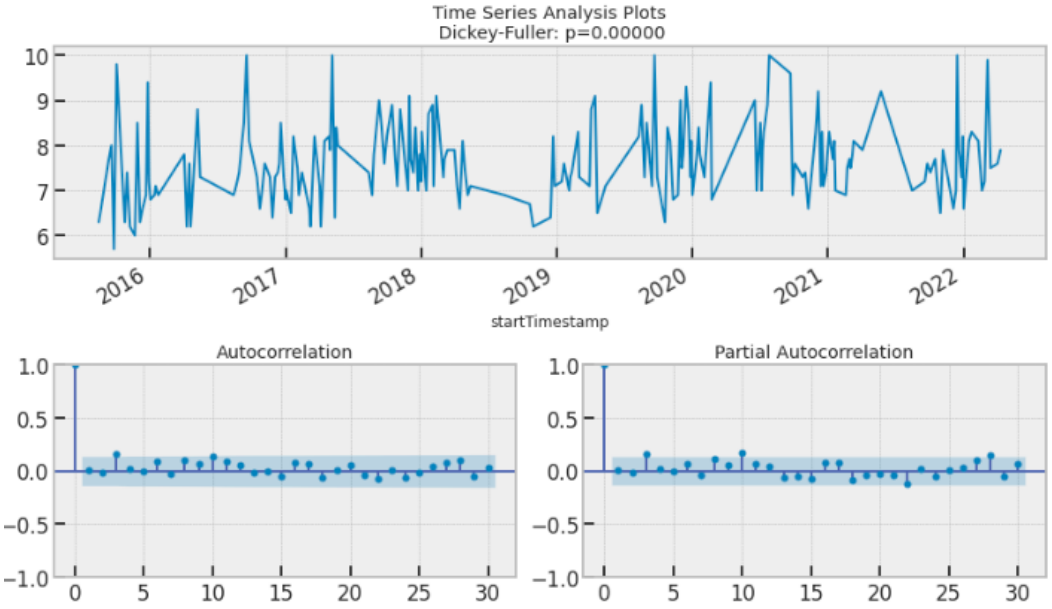


Figure 6.2. *Time series and autocorrelation plots of De Bruyne's performance*

The first graph shows the raw ratings through 7 years and the other two depict the correlation between values that are 30 time periods(lags) apart. From these plots, we see that values for the ACF are within 95% confidence interval (represented by the solid gray line) for lags > 0, which verifies that our data doesn't have any autocorrelation. We also notice that there is no seasonality in raw data. Zero p-values in the Dickey-Fuller test means that we reject the null hypothesis of presence of a unit root.

Same behavior is observed for all other players, although they are not presented for the sake of simplicity. These plots could indicate the value of terms to use on ARIMA, although we initially observe them to see how many timestamps in the past we want to look for relationships to current timestamp.

Summarizing, in the current problem we will correlate each game information with the previous one, creating a dataset of t-1 and t variables. Table 6.1 describes the features

that exist on every game(sample) in the dataset.

Table 6.1. *Features for every match*

Variable	Time
age	$t-1$
fifa rating	$t-1$
fifa potential	$t-1$
after injury	$t-1$
injury days	$t-1$
rest days	$t-1$
current team category	$t-1$
opponent category	$t-1$
home fixture	$t-1$
Performance	$t-1$
age	t
fifa rating	t
fifa potential	t
after injury	t
injury days	t
rest days	t
current team category	t
opponent category	t
home fixture	t
Performance	t

So, input data for a game of a player contain 10 variables that correlate with previous game and 9 variables that are known before next game and the purpose of the following approaches is to predict the 20th variable (performance) at time t and further. Finally, the experiments go one step further and try to forecast the performance of these soccer players on the next 10 games, namely $t+9$ steps ahead. To clarify this point, from approximately 200 games of every player, we cut out the 10 last games of February, March and April of 2022 and try to forecast their individual performance.

- Train set: N-10 games from August 2015
- Test set: 10 last games of 2022

Data as presented on figure 6.1 are in different scales. Most algorithms work better when data are transformed within specific scales. The methods with which we use standard scaling for the Forecasting are Linear regression, Support vector regression, Random Forest, XGBoost and MLP regressor. This type of scaling was described on section 5.3.2. LSTM utilizes Min Max Scaling, a method that transforms each feature to a given range, from 0 to 1 in this case.

6.2 Design and Methods

6.2.1 Modeling

On section 3.1, a detailed description of the seven learning algorithms was provided. One more model that is developed is a naive statistical model that predicts always the constant mean of all the historical performances of every player. The use of this model is intended to check if the learning models overpass its metrics and predictions. These seven methods are:

- Naive statistical mean
- Linear regression
- Support vector machines regression (SVR)
- Extreme gradient boosting regression (XGBoost)
- Random Forest regression (RF)
- Autoregressive integrated moving average (ARIMA)
- Multi layer perceptron (MLP)
- Long short-term memory (LSTM)

The reason that these methods were selected, is to cover a broad range of techniques. From simple statistical learning, to more complicated deep learning architectures. Some of these methods have a wide range of parameters that can be tuned and give promising results on model predictions.

Concerning model parameters, there was applied tuning on the hyperparameters through exhaustive search of the best ones. This search combines all specified parameter values for an estimator and retains the best combination of them. The decision for the best parameter grid is the smallest error or the biggest r^2 coefficient.

The process of the ARIMA is generally the same with the difference that in order to find the best model, we optimize for a given information criterion (AIC in our case). Except the response variable y, we use exogenous variables creating an ARIMAX model. These variables are used as additional features in the regression operation and do not include a constant or trend. This model can be regarded as a merger of the regression model (from exogenous) and the ARIMA model, as we described on section 3.1.5.

Random Forest, XGBoost and neural networks have more complex attributes to take into account and optimize. The space of parameters that used for the models in the experiments are listed in table 6.2.

Table 6.2. *Parameter grid for the forecasting models*

Model	Tuning Parameters
Linear regression	-
SVM	kernel=['linear', 'poly', 'rbf', 'sigmoid'], C=[0.5, 0.7, 1], epsilon=[7e-2, 1e-1, 1.0]
ARIMA	p (number of time lags)=[0-7], d (order of first-differencing)=[0,1,2], q (order of moving-average)=[0-7]
Random Forest	max features=['auto', 'sqrt', 'log2'], min samples leaf=[1, 8, 15], min samples split=[2, 8, 14], n estimators=[60, 80, 100, 120]
XGBoost	learning rate=[0.3, 0.09, 0.03], max depth=[5, 6, 7, 8], min child weight=[1,5], subsample=[0.5, 1], colsample bytree=[0.5, 1]
MLP	hidden layer size=[(50,50,50), (50,100,50), (50,100), (100,50), (100,)], activation=['tanh', 'relu'], solver=['sgd', 'adam'], learning rate=['constant','adaptive'], alpha=[0.0001, 0.05, 0.09]
LSTM	layers=[(50,1), (50,100,1), (20,1,20,100,1)], activation=['tanh', 'relu'], solver=['sgd', 'adam'], learning rate=['constant','adaptive']

Since neural networks are more complicated and difficult in understanding, we should pay more attention on LSTM. Several combinations of hidden layers, units per layer, batch size, learning rate, and embedding sizes were tested. However, Figure 6.3 visualizes the layers of an example model used for each player.

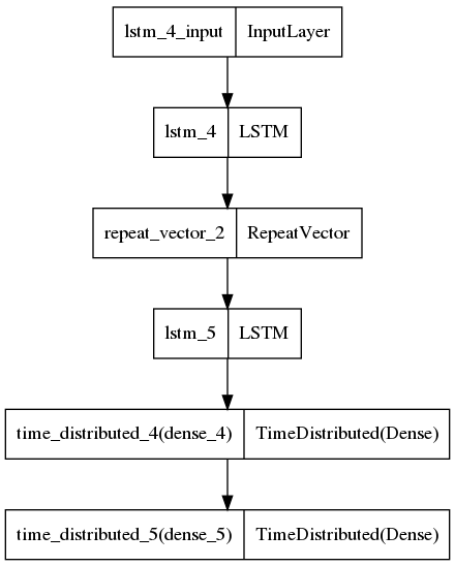


Figure 6.3. *LSTM architecture*

At first, the input is passed to 20 LSTM layers, then a vector repeats the input 1 time before 20 more LSTM cells. Then, a time distributed layer is applied to every temporal slice of the input with a dense activation of 100 layers and then the same two components giving an output of size 1. The process of designing, optimizing and fine tuning deep neural networks of this kind, is considered as art. It can be a long process, with a lot of trial and errors. However, on this dissertation this part is kept simple, since the optimal

architecture is not the main goal. The most proper way to validate that the network design was good is to see training and validation loss graph, see figure 6.4. Here, loss values on y-axis are calculated by taking the sum, of errors (predicted and real value) for each sample in the two sets.

Figure 6.4. *LSTM training and validation loss for N'Golo Kante*

If the curve decreases to a point of stability with a small difference between the two final loss values indicate a good match. Almost invariably, the training dataset will have a lower model loss than the validation dataset. Accordingly, there will likely be a discrepancy between the train and validation loss learning curves. The "generalization gap" is the name given to this discrepancy. A learning curve map indicates a good match if the training loss map flattens out at a certain point and a tiny gap exists between the validation loss plot and the training loss plot as it lowers to a stable point.

6.2.2 Software implementation

The process of creating a model for each one of the 30 time series is long and time demanding. Going one step further and search exhaustively for the best parameters for each model adds more implications to the development process. So, it's necessary to automate the code that develops all the steps described above.

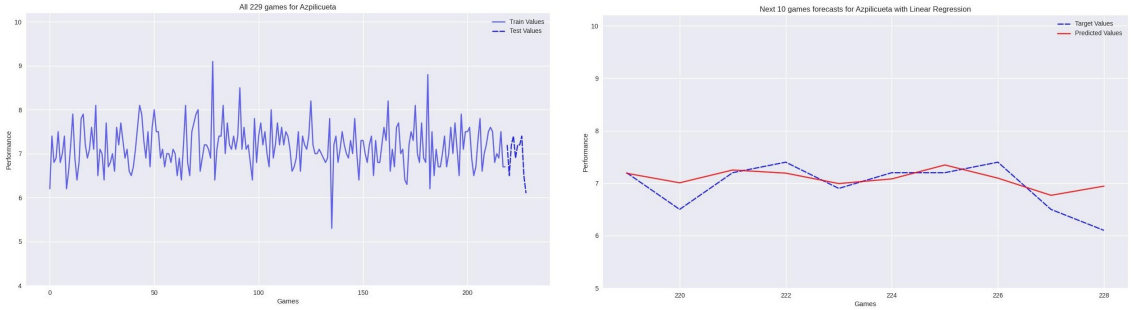
for each player in the dataset **do**
    Isolate his games to create a smaller dataset
    input-size ← 1
    output-size ← 1
    test-size ← 10
    Preprocess input
    Define model and run hyperparameter tuning
    Train model on train set with the best parameters
    Make predictions on test set
    Get metrics
    Plot real values and predictions
    Save all results
end for

The execution time for all the players of the dataset is about 2 hours. Statistical and machine learning methods complete faster than neural networks, as it is widely known, although on this problem the data are not so big and the time load is added by the parameter grid search. Extensive comments about execution time and optimization are omitted, since this isn't the main goal of this thesis.

6.3 Results and Discussion

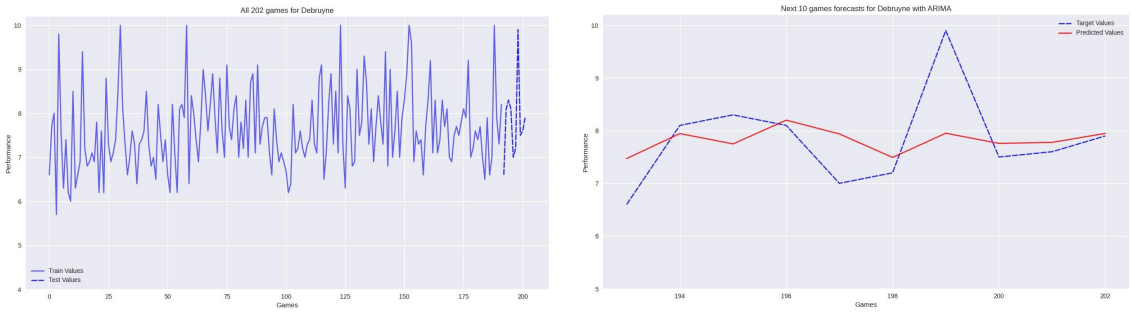
6.3.1 Visual exploration

After describing data attributes and methodologies, in this section we present the results of the experiments. We begin by visualizing the time-series of performance with the real values, the target values and the predictions. Since the whole dataset consists of 30 players and their corresponding time series, it would take a lot of pages to show all the forecasting plots. For this reason, it is interesting to examine one player for each of the three field-playing positions, attack, midfield and defense. Figures 6.5a, 6.6a and 6.7a show the progress of dependent variable 'performance' through the years and highlight the split (dashed line) of the last 10 games to forecast for each players. On the right, figures 6.5b, 6.6b and 6.7b, present these cut out games that are considered as next and whose rating values the models try to predict.



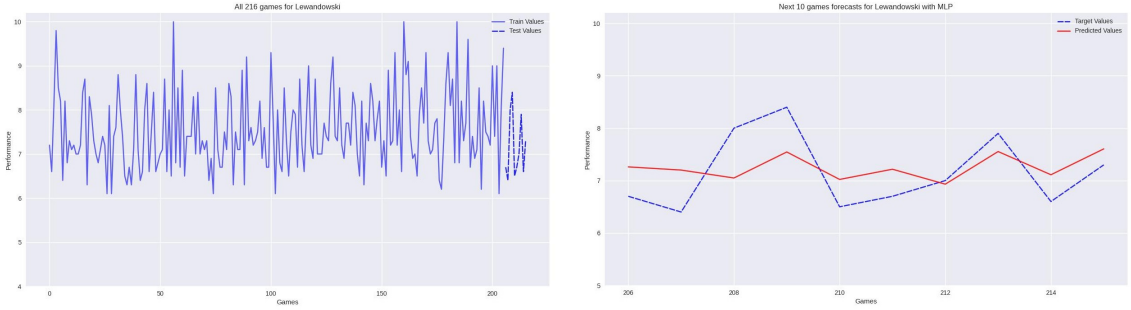
(a) *Train test split from real values* **(b)** *Target values on test set vs predictions*

Figure 6.5. *Real values and forecasts for Azpilicueta*



(a) *Train test split from real values* **(b)** *Target values on test set vs predictions*

Figure 6.6. *Real values and forecasts for De Bruyne*



(a) *Train test split from real values* **(b)** *Target values on test set vs predictions*

Figure 6.7. *Real values and forecasts for Lewandowski*

Chelsea's defender, Cezar Azpilicueta, has a smooth performance history with very low variance, figure 6.5. The best model for him is linear regression, which predicts quite successfully the next 10 games. The only exception may seem the 10th match where the gap between target and prediction is bigger. However, the general behavior of the dependent variable is followed very closely and we have a very clear view of how good appearances he is about to make.

Subsequently, Manchester City's central midfielder, Kevin De Bruyne (figure 6.6), is considered by many the best play maker in the world and a constant threat for the opponents. This fact is indicated by his high mean and some extraordinary raises with perfect ratings of 10. The performance of a footballer with such a high variance is difficult to be forecasted. Contrariwise, ARIMA, has very good results and predicts very closely his performance on the next 10 games. Of course, sudden raises in performance, which have the nature of outliers, cannot be approached very good from the dependent and the exogenous attributes. Although, almost all the increases and decreases, along with this spike in the 7th place are predicted nicely, but more smoothly.

Lastly, one of the best strikers in the world these years, Bayern Munich's center forward, Robert Lewandowski (figure 6.7), has a similar time series plot with De Bruyne, with a lot of sudden raises and falls, in contrast with Azpilicueta's more constant performances. MLP has the best results on this player, revealing very clearly the behavior of the player in the future games. For example, the values in the 3rd and 4th match (208,209) of the test set are much higher than the predicted ones, but the increase in his performance is showed correctly. The difficult pattern of Lewandowski is followed not identically, but very well and the differences in performance are forecasted correctly, though in a smoother way.

6.3.2 Comparison and evaluation

A good test to examine that the patterns learned from the dataset are important and have substance is to create a benchmark that compares the learning techniques with random or naive methods. If the learning models have worse predictions and therefore metrics than the plain methods then there is no point on inserting such complexity on this type of data. Table 6.3 presents the lowest Mean Absolute Error(MAE) on test set

(next 10 games) for each player in the dataset. Here, we check if the naive statistical model, which uses the historical mean for predictions, is better than all the other models.

Table 6.3. *Lowest MAE per player*

Player	Best model	MAE	Player	Best model	MAE
Messi	Linear Reg	0.621	Neymar	Random Forest	0.798
Lewandowski	MLP	0.727	Suarez	Random Forest	0.384
Jorginho	Linear Reg	0.406	Kjaer	MLP	0.228
Benzema	Linear Reg	0.734	Mount	Linear Reg	0.769
Kante	XGBoost	0.333	Mahrez	MLP	0.533
Cristiano	Random Forest	0.573	Bruno	ARIMA	0.569
Salah	Linear Reg	0.640	Lautaro	MLP	0.587
Debruyne	Linear Reg	0.503	Kane	XGBoost	0.601
Mbappe	LSTM	0.932	Pedri	Random Forest	0.561
Donnarumma	Random Forest	0.327	Foden	Naive	0.253
Haaland	Naive	0.657	Moreno	MLP	0.922
Lukaku	MLP	0.502	Barella	MLP	0.472
Chiellini	XGBoost	0.290	Dias	Naive	0.210
Bonucci	XGBoost	0.440	Modric	ARIMA	0.459
Sterling	XGBoost	0.760	Azpilicueta	Linear Reg	0.255

There are some interesting results in the table of Mean Absolute Errors. Firstly, there is a variety of the best performing models for each player with statistical, linear and tree models sharing frequencies with neural networks. Secondly, the average MAE for all players is around 0.53. Forecasting values half a unit up or down is quite satisfying considering that the most players in the dataset are in the attacking part of the field and their performances are not so stable containing spikes.

The most important element of table 6.3 although is that the Naive model overpass all learning models only on three cases, for Haaland, Foden and Ruben Dias. The common fact for the three players is that have played very few games, see figure 4.3. The short length of their datasets do not help the machine learning techniques to find behaviors and patterns in the features. Ruben Dias, may have more games than the other two, although his performance is very stable in the next 10 games and that's why the naive model that infers historical average as predictions has better results.

Finally, the fact that the performance of 27 from the 30 total players is predicted better with learning algorithms, means that there is valuable information in the created dataset and these techniques have substance.

Another metric to evaluate forecasting algorithms is the Root Mean Squared Error. Its purpose is approximately the same with MAE, namely to minimize this error. Table 6.4 gathers all RMSEs for the separate players of the dataset. Like MAE, this error is calculated on the test set of the next 10 games for each player.

Again, there is a variety on the algorithms that produce the best results for each

Table 6.4. *Lowest RMSE per player*

Player	Best model	RMSE	Player	Best model	RMSE
Messi	Linear Reg	0.712	Neymar	MLP	1.129
Lewandowski	MLP	0.783	Suarez	Random Forest	0.486
Jorginho	ARIMA	0.615	Kjaer	MLP	0.302
Benzema	SVM	0.855	Mount	Linear Reg	0.921
Kante	XGBoost	0.417	Mahrez	MLP	0.665
Cristiano	XGBoost	0.890	Bruno	Linear Reg	0.635
Salah	Linear Reg	0.753	Lautaro	MLP	0.885
Debruyne	ARIMA	0.772	Kane	XGBoost	0.765
Mbappe	LSTM	1.170	Pedri	Linear Reg	0.679
Donnarumma	LSTM	0.342	Foden	Naive	0.310
Haaland	Naive	0.746	Moreno	Linear Reg	1.162
Lukaku	MLP	0.563	Barella	MLP	0.677
Chiellini	Naive	0.342	Dias	Naive	0.282
Bonucci	Random Forest	0.646	Modric	ARIMA	0.564
Sterling	Linear Reg	0.942	Azpilicueta	Linear Reg	0.350

player. All models are met at least one time in RMSEs table (6.5). The average RMSE for all players is very low, 0.67. However, the difference here is that the plain(Naive) model is the best not only on the three aforementioned footballers. Naive RMSE is the lowest also for Juventus captain, Giorgio Chiellini. A good explanation is that like Ruben Dias has constant performance on the next 10 games, so predicting the average as a rating is a good solution. Moreover, he has few games(131) in his dataset -figure 4.3- due to the severe injuries he suffered from these 7 years. Lastly and most importantly, the truth that learning algorithms have the lowest RMSEs on performance forecasting for 26 out of the total 30 players indicates that the dataset was built with valuable information and that these methods are effective.

The third comparison used to benchmark the forecasting methods is to check the performance directions. Since the deviation on the dependent variable of our dataset is very big, as explained before in this dissertation we would expect the forecasts to be smoother than the sudden large or small real values. Although smoother, the forecasting could find the improvement or fall in performance from the last game. If these ups and downs are better that randomness, the models would be successful. The best correct predicted directions between matches, on the unseen test set again, are presented in table 6.5.

Like previous tables on ths section, all models appear at least one time. It should be mentioned that Naive model absents since the constant predicted mean has not ups and downs. Another thing that must be clarified, is that three directions were taken into account. Directions labeled as 1 meant increase in performance, label 0 meant decrease and label 2 indicated exactly the same performance with the previous match.

We observe that all correct directions reach at least the levels of randomness, which

Table 6.5. *Correct predicted directions*

Player	Best model	Directions	Player	Best model	Directions
Messi	Linear Reg	8/10	Neymar	SVM	6/10
Lewandowski	LSTM	8/10	Suarez	SVM	5/10
Jorginho	Random Forest	8/10	Kjaer	SVM	7/10
Benzema	Linear Reg	7/10	Mount	Linear Reg	6/10
Kante	XGBoost	8/10	Mahrez	SVM	8/10
Cristiano	XGBoost	7/10	Bruno	XGBoost	8/10
Salah	LSTM	5/10	Lautaro	Linear Reg	6/10
Debruyne	Linear Reg	7/10	Kane	ARIMA	7/10
Mbappe	Random Forest	9/10	Pedri	XGBoost	6/10
Donnarumma	XGBoost	6/10	Foden	ARIMA	8/10
Haaland	Linear Reg	8/10	Moreno	MLP	6/10
Lukaku	SVM	5/10	Barella	Linear Reg	7/10
Chiellini	Random Forest	7/10	Dias	MLP	9/10
Bonucci	MLP	8/10	Modric	ARIMA	9/10
Sterling	Linear Reg	7/10	Azpilicueta	MLP	8/10

is 5/10. Only three footballers, Salah, Lukaku and Suarez, are on this random level. Here, the average correct predicted directions for all players is 7.2 out of 10, a number that overpass by far the randomness level of 5. Dias, Modric and Mbappe have the higher number(9/10) of successful predicted directions, with their forecasting failing only on one game.

Summarizing, the fact that the predictive modeling achieves better results than randomness on 27 of the total 30 elite soccer players is very encouraging and allows us to have a clear view of the performance progress of these footballers in the near future.

Chapter 7

Conclusion and future work

This thesis made a contribution to the field of sports analytics by researching how predictive modeling is applied to future individual performance. The contributions of this thesis are enumerated below, along with suggestions for further investigation.

7.1 Conclusions

It is clear that sports analytics will play a significant role in how well a club performs in the upcoming years. There will be a ton of data, and with the right methods of exploitation, it can be used to predict important factors of the game.

The exploration of a single, untapped area—the use of predictive modeling for future individual performance of soccer players—represents the thesis' overall contribution to sports analytics. Having a view of the performance progress in the near or distant future can have a significant effect on the individual as well as the team on many different levels. The thesis also discussed the significance of data collection and the standardization of data storage and processing in football while examining the efficacy of various approaches and algorithms for specific challenges. The contributions to each investigation are more thoroughly described below:

- Firstly, there was made an attempt to create a solid dataset of soccer players, since there are not publicly available data in the kind that suit the problem formulation. The group of players was limited on elite athletes and along with their ratings, more features like injuries, FIFA skills, club values and more were gathered or created. Player attributes, pre-game information, team details and post-game information can affect performance of individuals in a soccer game.

- Secondly, several regression techniques attempted to predict the performance in one next match, using pre-game information and previous performances. Additionally, it was studied which attributes affect player's performance and generalize the results of this small sample of 30 footballers to the whole population of professional players. It was indicated that some features like player's form, opponent's difficulty and rest days, are more important than others in different ways and that even simple methods like linear regression provide reasonably good performance.

- Finally, the main contribution of this thesis is the forecasting of performance of individual soccer players in one or more specific future matches. Various statistical, machine learning and deep learning methods were applied to each separate player's time series of historical performances and other explanatory variables and were compared with random predictions and naive models. The performance of each player in future games is predicted very well and learning algorithms' results are a lot better than random and naive predictors. Therefore, the sports professional can use these techniques and have a clear view of how good or bad appearances his player is about to make in the near or distant future.

7.2 Future Directions

Given the limited amount of research in that field, this thesis establishes the parameters within which further investigation might be expanded. Future research could go in a number of different areas. We'll go through some broad recommendations in this section.

Firstly, additional data may be discovered by future research and made available to the current models. It was discussed in Chapter 4 that gathering data on soccer is complex and challenging. In each study, this thesis succeeded in demonstrating that it is possible to create performance-predictive models for soccer players even with little data. However, more data would significantly improve the performance of any model.

More precisely, the inclusion of data from clubs would be advantageous for all research. Training information like exposure records, coach reports as well as the GPS data, could be a good starting point. Event data like team playing styles and individual actions like shots, passes, air duels etc could also bring value to the modeling. Player similarities, teammate interactions, manager reports, updated market values and betting odds for future games are variables that could reveal better patterns on individual performance. A last feature to consider is psychological situation and out-of-field life of footballers, since human beings are so complicated entities and factors from one aspect of life affect other fields in a good or a bad way.

Secondly, for each of the investigations described in this thesis, a wider range of models can be explored. Although the appropriateness of different models was explored in this research, other techniques could be applied. Obviously, the scope of this thesis did not allow for a comprehensive comparison of all regression or forecasting algorithms. The selection of the present methods was driven by a variety of considerations, including an algorithm's applicability to a given situation and its performance in prior challenges with analogous traits. But there is a lot of research being done in machine learning, and new algorithms are constantly being released. Therefore, there is still need for research into the approaches that work the best for a given issue.

Thirdly, additional research on the studies is necessary to see how broadly the findings may be applied. This specific research were carried out for a particular group of 30 top class players. However, since there can be characteristics, ranging from elite club player treatments to individual attributes, that may be unique to this players and teams, it is still unclear whether these studies can translate effectively to other situations.

It is worth looking into whether the specific models that were selected are also the best for jobs involving performance forecasting across all leagues and divisions. To draw firm conclusions on these issues, more investigation is necessary, with players from various nations and clubs, and most importantly players not being the best in the world.

Bibliography

[1] Arndt, C., and Brefeld, U. Predicting the future performance of soccer players. *Statistical Analysis and Data Mining 9*, 5 (Oct. 2016), 373–382.

[2] Arosha Senanayake, S. M. N., Malik, O. A., Iskandar, P. M., and Zaheer, D. A knowledge-based intelligent framework for anterior cruciate ligament rehabilitation monitoring. *Applied Soft Computing 20* (July 2014), 127–141.

[3] Bartolucci, F., and Murphy, T. B. A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports 11*, 4 (Dec. 2015), 193–203. Publisher: De Gruyter.

[4] Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (New York, NY, USA, July 1992), COLT '92, Association for Computing Machinery, pp. 144–152.

[5] Breiman, L. Random Forests. *Machine Learning 45*, 1 (Oct. 2001), 5–32.

[6] Brooks, J., Kerr, M., and Guttag, J. Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, Aug. 2016), KDD '16, Association for Computing Machinery, pp. 49–55.

[7] Chen, T., and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, Aug. 2016), KDD '16, Association for Computing Machinery, pp. 785–794.

[8] Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Machado Pereira, A. C., and Nassis, G. P. Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. *Sports Medicine - Open 5*, 1 (July 2019), 28.

[9] Constantinou, A. C., Fenton, N. E., and Neil, M. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems 36* (Dec. 2012), 322–339.

[10] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20*, 3 (Sept. 1995), 273–297.

[11] De Mauro, A., Greco, M., and Grimaldi, M. A formal definition of Big Data based on its essential features. *Library Review 65* (Mar. 2016), 122–135.

[12] Diana, B., Zurloni, V., Elia, M., Cavalera, C. M., Jonsson, G. K., and Anguera, M. T. How Game Location Affects Soccer Performance: T-Pattern Analysis of Attack Actions in Home and Away Matches. *Frontiers in Psychology 8* (2017).

[13] Dixon, M. J., and Coles, S. G. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 46*, 2 (1997), 265–280.

[14] Elmiligi, H., and Saad, S. Predicting the Outcome of Soccer Matches Using Machine Learning and Statistical Analysis. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (Jan. 2022), pp. 1–8.

[15] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. Knowledge Discovery in Databases: An Overview. *AI Magazine 13*, 3 (Sept. 1992), 57–57. Number: 3.

[16] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 5 (Oct. 2001), 1189–1232.

[17] Friedman, J. H. Stochastic gradient boosting. *Computational Statistics & Data Analysis 38*, 4 (Feb. 2002), 367–378.

[18] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. Time Series Analysis: Forecasting and Control, 5th Edition | Wiley, 2015.

[19] Gholamy, A., Kreinovich, V., and Kosheleva, O. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)* (Feb. 2018).

[20] Goddard, J. Regression models for forecasting goals and results in professional football. *International Journal of Forecasting 21* (Apr. 2005), 331–340.

[21] Groll, A., and Abedieh, J. Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports 9* (Jan. 2013), 51–66.

[22] Gréhaigne, J.-F., Griffin, L. L., and Richard, J.-F. *Teaching and Learning Team Sports and Games*. Psychology Press, 2005.

[23] Haase, J., and Brefeld, U. Finding Similar Movements in Positional Data Streams. In *Machine Learning and Data Mining for Sports Analytics@Principles and Practice of Knowledge Discovery in Databases/ECML* (2013), pp. 49–57.

[24] Hand, D. J. Principles of Data Mining. *Drug Safety 30*, 7 (July 2007), 621–622.

[25] Henderson, G., Barnes, C., and Portas, M. Factors associated with increased propensity for hamstring injury in English Premier League soccer players. *Journal of Science and Medicine in Sport 13* (July 2010), 397–402.

[26] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (Nov. 1997), 1735–1780.

[27] Hughes, M. D., and Bartlett, R. M. The use of performance indicators in performance analysis. *Journal of Sports Sciences* (Dec. 2010), 739–754.

[28] Hyndman, R. J., and Athanasopoulos, G. *Forecasting: Principles and Practice (2nd ed)*. 2018.

[29] Hyndman, R. J., and Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting 22*, 4 (Oct. 2006), 679–688.

[30] Jae Sik, L. Prediction of pitch type and location in baseball using ensemble model of deep neural networks. *Journal of Sports Analytics* (Feb. 2022), 1–12.

[31] Kharrat, T., McHale, I. G., and Peña, J. L. Plus–minus player ratings for soccer. *European Journal of Operational Research 283*, 2 (June 2020), 726–736.

[32] Kohavi, R., and Wolpert, D. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* (San Francisco, CA, USA, July 1996), ICML'96, Morgan Kaufmann Publishers Inc., pp. 275–283.

[33] Kolen, J. F., and Kremer, S. C. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In *A Field Guide to Dynamical Recurrent Networks*. IEEE, 2001, pp. 237–243. Conference Name: A Field Guide to Dynamical Recurrent Networks.

[34] Kotu, V., and Deshpande, B. Chapter 12 - Time Series Forecasting. In *Data Science (Second Edition)*, V. Kotu and B. Deshpande, Eds. Morgan Kaufmann, Jan. 2019, pp. 395–445.

[35] Lames, M., and McGarry, T. On the search for reliable performance indicators in game sports. *International Journal of Performance Analysis in Sport 7*, 1 (Jan. 2007), 62–79.

[36] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521*, 7553 (May 2015), 436–444. Number: 7553 Publisher: Nature Publishing Group.

[37] Matteo, D., Gastin, P., Suppiah, H., and Carey, D. Predicting Athlete Performance in Team Sports Using Nearest Neighbour Modelling. In *Proceedings of the 9th International Performance Analysis Workshop and Conference & 5th IACSS Conference* (Cham, 2022), A. Baca, J. Exel, M. Lames, N. James, and N. Parmar, Eds., Advances in Intelligent Systems and Computing, Springer International Publishing, pp. 101–108.

[38] Medsker, L., and Jain, L. C. *Recurrent Neural Networks: Design and Applications*. CRC Press, Dec. 1999.

[39] Michael Stöckl, Thomas Seidl, Daniel Marley, and Paul Power. Making Offensive Play Predictable -Using a Graph Convolutional Network to Understand Defensive Performance in Soccer. In *Proceedings of the 15th MIT Sloan Sports Analytics Conference* (2021), vol. 2022.

[40] Molodchik, M., Paklina, S., and Parshakov, P. Peer Effects on Individual Performance in a Team Sport. *Journal of Sports Economics 22* (June 2021), 571–586.

[41] Moroney, M. J. *Facts from figures*. Penguin Books, Harmondsworth [England]; Baltimore, 1956. OCLC: 14955379.

[42] Pantzalis, V. C., and Tjortjis, C. Sports Analytics for Football League Table and Player Performance Prediction. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA* (July 2020), pp. 1–8.

[43] Pappalardo Luca, Cintia Paolo, Ferragina Paolo, Massucco Emanuele, Pedreschi Dino, and Giannotti Fosca. PlayeRank. *ACM Transactions on Intelligent Systems and Technology (TIST)* (Sept. 2019). Publisher: ACM PUB27 New York, NY, USA.

[44] Pariath, R., Shah, S., Surve, A., and Mittal, J. Player Performance Prediction in Football Game. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (Mar. 2018), pp. 1148–1153.

[45] Porter, J. Predictive Analytics for Fantasy Football: Predicting Player Performance Across the NFL. *Honors Theses and Capstones* (Jan. 2018).

[46] Redwood-Brown, A., Bussell, C., and Bharaj, H. The impact of different standards of opponents on observed player performance in the English Premier League. *Journal of Human Sport and Exercise 7* (June 2012).

[47] Reep, C., Pollard, R., and Benjamin, B. Skill and Chance in Ball Games. *Journal of the Royal Statistical Society. Series A (General) 134*, 4 (1971), 623–629.

[48] Richard J. Brook, and Gregory C. Arnold. Applied Regression Analysis and Experimental Design, 2018. Publisher: CRC Press.

[49] Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., and Medina, D. Effective injury forecasting in soccer with GPS training data and machine learning. *PLOS ONE 13*, 7 (July 2018), e0201264.

[50] Schultze, S., and Wellbrock, C.-M. A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics 4*, 2 (Oct. 2017), 121–131.

[51] Shah, R., and Romijnders, R. Applying Deep Learning to Basketball Trajectories. *arXiv preprint arXiv:1608.03793* (2016).

[52] Sharma, M., Srivastava, R., Anand, A., Prakash, D., and Kaligounder, L. Wearable motion sensor based phasic analysis of tennis serve for performance feedback. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), 5945–5949.

[53] Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, G., Grossniklaus, M., and Keim, D. A. Bring It to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 13–22. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[54] Tom Mitchell. The Discipline of Machine Learning. Publisher: Carnegie Mellon University, School of Computer Science, Machine Learning.

[55] Wright, M. 50 years of OR in sport. *Journal of the Operational Research Society 60* (Feb. 2009). S161-S168.