Player Evaluation Using Machine Learning

by

Jak Cullinane

This thesis has been submitted in partial fulfillment for the degree of Bachelor of Science in Software Development

in the
Faculty of Engineering and Science
Department of Computer Science

April 2024

Declaration of Authorship

This report, Player Evaluation Using Machine Learning, is submitted in partial fulfillment of the requirements of Bachelor of Science in Software Development at Munster Technological University Cork. I, Jak Cullinane, declare that this thesis titled, Player Evaluation Using Machine Learning and the work represents substantially the result of my own work except where explicitly indicated in the text. This report may be freely copied and distributed provided the source is explicitly acknowledged. I confirm that:

- This work was done wholly or mainly while in candidature Bachelor of Science in Software Development at Munster Technological University Cork.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Munster Technological University Cork or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Jak Cullinane

Date: 26/04/2024

MUNSTER TECHNOLOGICAL UNIVERSITY CORK

Abstract

Faculty of Engineering and Science
Department of Computer Science

Bachelor of Science

by Jak Cullinane

Football is the most popular sport in the world. Every movement within the sport causes a gigantic ripple effect across the entirety of the fan base. With the growth of football in terms of finances and influence, transfers between clubs continue to grow in importance. With the sports evolution and growing use of data analytics, attempts are being made in order to predict

This project aims to measure a football player's ability and predict which league they would be most suited for. It is essential to understand and evaluate the factors contributing to a player's performance and establish a method for predicting a league most suited to the player.

Acknowledgements

I would like to acknowledge and thank Dr. Alex Vakaloudis for his assistance and guidance throughout the duration of the project, providing me with the ability and support to research and develop this project.

I am grateful and thankful to Munster Technological University for providing me with the tools and knowledge in order to tackle this project.

Finally, I would like to thank my friends and family for providing me with the support and reassurance to complete this project.

Contents

List of Figures

Abbreviations

CPU	**C**entral **P**rocessing **U**nit
JOI	**J**oint **O**fensive **I**mpact
JDI	**J**oint **D**efensive **I**mpact
VAEP	**V**aluing **A**ctions **E**stimating by **P**robabilities
xG	Expected **G**oals
xGot	Expected **G**oals **O**n **T**arget
xA	Expected **A**ssits
xP	Expected **P**ass completion
API	**A**pplication **P**rogramming **I**nterfaces
KPI	**K**ey **P**erformance **I**ndicators
ANOVA	**AN**alysis **O**f **VA**riance
CM	**C**enti**M**eters
CSV	**C**omma **S**eparated **V**alues
OLS	**O**rdinary **L**east **S**quares
GBR	**G**radient **B**oosted **R**egressor
EN	**E**lastic **N**et
KNN	**K** **N**earest **N**eighbours
RFR	**R**andom **F**orest **R**egression
DTR	**D**ecision **T**ree **R**egression
MAE	**M**ean **A**bsolute **V**ariance
R - Squared	Coefficient of Determination
RMSE	**R**oot **M**ean **S**quare **D**eviation

Chapter 1

Introduction

1.1 Motivation

Football is the most popular sport in the world. Every movement within the sport causes a gigantic ripple effect across the entirety of the fan base. With the growth of football in terms of finances and influence, transfers between clubs continue to grow in importance. With the sports evolution and growing use of data analytics, attempts are being made in order to predict

This project aims to examine the challenges football players encounter across various leagues and to ascertain the feasibility of developing a model for predicting their compatibility with a given league.

1.2 Contribution

This project focuses on machine learning and data analytics. During my studies, my classes, such as Machine Learning, Probability and Statistics, and Programming for Data Analytics, have helped immensely with the skill set to tackle this project. Skills such as comprehending and manipulating statistics to understand and utilize them within the project, necessary experience with Python and machine learning libraries to lay the foundation for the project, and data analytical skills are vital to the project. These, in tandem, have provided me with the foundational knowledge to approach this project from a technical point of view while still providing a path for growth throughout the project by attempting to maximize my skill set and take on more significant challenges. Furthermore, classes such as Agile Processes and the Group Project have provided a basis for project management skills so I can guide the project's development throughout the project.

1.3 Structure of This Document

An explanation of the structure of the report and a brief introduction to each chapter.

- Chapter 1 is an introduction to the project, providing information on the motivation, contribution and the structure.
- Chapter 2 provides an understanding of the thematic area within Computer Science and a Literature review of predictive analysis of football players' performance to determine the most suitable league for them.
- Chapter 3 is an explanation of the problem this project attempts to address and provides a solution for.
- Chapter 4 this section will lay out the plan to implement a solution for the problem identified and addressed in chapter 3.
- Chapter 7 a discussion on conclusions from the project and future work.

Chapter 2

Background

2.1 Thematic Area within Computer Science

1. Core Topic:
   - Predictive analysis of football players' performance to determine the most suitable league for them.
2. Core Area(s):
   - Data analytics
   - Sport analytics
   - Machine learning
3. Main Area(s) within Computer Science:
   - Data Science
   - Artificial Intelligence

2.2 Literature Review

This project aims to measure a football player's ability and predict which league they would be most suited for. It is essential to understand and evaluate the factors contributing to a player's performance and establish a method for predicting a league most suited to the player. In this section, I will review the literature on each area to establish a frame of reference for the project.

2.2.1 Conditions that affect performance

Performance in sports is a broad term and encompasses many things, such as the match outcome, scoring contributions, defensive contributions, and team interactions. For this paper, I define performance as ”the accomplishment of goals by meeting or exceeding predefined standards” [1]. When viewing an athlete’s performance, we must first look at what can cause that performance. Various factors can influence an athlete’s performance, including ”physical power, mental strength, and mechanical edge” [2]. However, there are equally important influences such as recovery, cognitive ability, emotional regulation, exterior influences, and team chemistry. This section will investigate these attributes and the overall importance of performance in sports.

2.2.2 Physiology

Physiology, the study of how a body functions, is undoubtedly one of the most influential aspects of performance. More importantly and accurately, the discipline of exercise or sports physiology, which has derived from its parent physiology, works to study ”how the body’s functions are altered when we are physically active” [3]. This will include the response to strenuous activity and how fitness, endurance, and body composition interact with physiology.

There is a direct relation between a footballer’s fitness and quality. This makes sense because fitness represents the ability to operate at a high output level for a more extended period of time. This allows footballers to operate at their peak technical ability for as long as possible, which means that their influence on a game will be more significant than that of less-fit players. It was found that high-tier and moderate footballers perform a similar amount of standing, walking, and running at a low intensity. However, high-tier footballers will perform more high-intensity running, sprinting, and running backward [4].

Most athletes will have the genetics best suited to their sport at the highest level of competitive sport. Furthermore, they will undergo consistent and intense training to a high standard overlooked by and managed by the best coaches in their field. However, is it worth acknowledging that some athletes stand beyond everyone else due to a combination of dedication to training and their genetic advantages, such as Michael Jordan, Lebron James, Michael Phelps, Usain Bolt, and Cristiano Ronaldo, who are known for dominating their respective sports and represent absolute athleticism in their respective sport? For instance, typically, swimmers will have longer torsos and shorter legs. Michael Phelps, who is 6 feet 4 inches, has the lower body of a man 8 inches shorter and the torso of a man 4 inches taller [5].

FIGURE 2.1: Ronaldo jumping prowess [6]

2.2.3 Recovery

Recovery is an essential part of the process that allows players to perform at the peak of their abilities. In order to properly recover, several areas must be attended to. These include sleep, rest, nutrition, and injury treatment. At the highest level of sport, an athlete's body's demands to perform at the highest level are beyond an average person's demands.

Nutrition is essential for athletes. Caloric intake will increase due to the energy expended during training and events. The necessary calories must be consumed along with the correct balance of "nutrients essential for the formation of new body tissues and proper functioning of the energy systems that work harder during exercise." these include carbohydrates, fats, vitamins, proteins, minerals, and water [2].

Sleep and rest are vital and a requirement for everyone, when competing at the highest level of a sport, the ability to exert yourself to maximum capability is essential. While we consider sleep to be one of the most critical elements of having a good day and being productive, many athletes appear to suffer from poor sleep. While there is evidence to suggest that sleep deprivation can have serious adverse effects, many athletes suffer from sleep restriction. The effects of sleep restriction are severe, and conflicting conclusions can not be drawn. However, the adverse effects of sleep loss on cognitive function [7] are apparent and concern athletes.

Injuries are a part of sport and can cause significant issues for athletes. Some injuries are so devastating that players never recover properly from them and perform at a lower level prior to the injury. Often, players will need to trust the body part in which the injury occurred, and their ability to reach the peak of their abilities is hampered. According to [8], physical stress, psychological stress, and recovery are essential for illness, and monitoring these indicators can provide insight into preventing injuries and illness for players.

2.2.4 Technical and Tactical ability

Technical ability is one of the most significant indicators of a player's overall ability. In a systematic review [9], 93% of studies showed that technical ability was crucial to talent identification. However, tactical ability is equally important as an indicator, and according to this study [10], there is no positive correlation between technical and tactical ability. This means these are distinctly different skills and must be treated as such when viewing a player. One could argue this is a manifestation of ”talent” and ”hard work,” which combine to create a high-tier athlete. Furthermore, the study breaks down ”offensive” and ”defensive” tactical performance and finds a low correlation between both. From this, the players' specialties and positions show themselves in the data. It is essential to recognize the position of a player when viewing their technical and tactical ability.

2.2.5 Emotion regulation

Emotions are closely tied when playing any sport or competing in any event. ”Emotions are a critical parameter to take into account for anyone trying to reach peak performance” [11] . One might feel many different emotions about why playing sports and the relation these emotions have to performance are essential. There is a positive increase in performance when anger is experienced in sports that require physical power [12]. Furthermore, we can see differences in how different personalities respond to emotions. The study expanded that extroverts benefited more from anger than introverts [12]. Some players can maintain their emotional condition by maintaining an ego.

Many top players have an ego about being the best. Zlatan Ibrahimovic is famous for this ego. Cristiano Ronaldo is known for his undying work ethic and motivation. Both players can use their unique characteristics as a pillar of strength, relying on it in times of difficulty. This is particularly important for players who play for teams competing for their sport's highest awards.

2.2.6 Exterior influences

Besides individual characteristics and abilities, other factors go into performance. Some of these include home advantage, pitch conditions, weather conditions, and travel for the game.

Significant data proves that home advantage is an actual influence on games. The effect of the crowd is potentially the most significant contributing benefit to playing at home [13]. Furthermore, the distance traveled for away games plays a significant factor. It was found that in-home derbies, the benefit of home and away is significantly reduced [13]. This is likely because both sides of fans can easily acquire tickets. Considering this, it is essential to contextualize games played by a player and check for any skewing towards home or away games. It would be expected that a player who plays most of their games away is actually performing at a higher level compared to those who play a split. However, consideration for derbies and cup competitions such as the Champions League should be made due to the nullifying effect of home advantage in derbies.

2.2.7 Team bonding

When operating within team sports, all players in a team contribute to a performance. Some of the most outstanding individual performances ever in sports are based on the foundation of a team that complements that player. While Michael Jordan is regarded as the greatest player of all time in basketball by many, his performances were greatly enhanced by his teammates Dennis Rodman and Scottie Pippen. Michael Jordan's teammates could dominate other areas of the game, providing assists and allowing him to find positions where he could be best utilized.

In a study on player chemistry within football teams, the metrics Joint Offensive Impact (JOI) and Joint Defensive Impact (JDI) were introduced in order to calculate and predict the chemistry of players [14]. These metrics are based on the Valuing Actions by Estimating Probabilities (VAEP) framework, which will be further analyzed in the next section. This paper was able to view players' defensive and offensive output and demonstrate the changes when paired with different partners. They were displaying the genuine effect of chemistry on players. This paper was fascinating as it went on to predict the chemistry of players that had never played before and the chemistry of a player within a team they had never played for.

2.2.8 Modern Scouting

Scouting in modern sports is becoming more data-focused. This means that all scouting is done through the lens of statistics rather than the old-fashioned ”eye test.” This ranges from simple goals and assists to more advanced metrics. However, old methods of relying on social relationships and trust built with individuals remain a large part of the process [15]. For the sake of this project, this section will look at the growing data-focus aspect of scouting.

2.2.9 Technology

Many technologies are used to track and monitor players and store the data collected. Video analysis, player tracking systems, and vast databases are used to store every piece of information. ChyronHego's TRACAB system is one of the leaders in optical tracking systems and is being used widely. What is called the ”Gen5” system can track ”spatio-temporal tracking variables” [16].

2.2.10 Foundational

Modern scouting continues to use a foundation for developing and scouting players. This will typically consist of young players recruited and playing for a developing team or academy. They will be monitored and tracked on their growth and expected to hit a foretasted potential from when they were scouted. Some clubs will have sister or feeder clubs where a relationship has been established. This is a mutually beneficial relationship where young, talented players can develop at this club, or talented players brought to the club may move to the bigger one. Data collection is at its best at the highest levels of any sport and in mainly developed countries such as the United States. However, talent is found worldwide, so a global scouting network is used. By creating an optimal environment to foster competitiveness and encourage and nurture talent, academies have positive outcomes for players and clubs. Academies lead to outcomes such as senior progression and player development. Many players come from various backgrounds, and the academy can provide a stable environment and a great equalizer for all developing players [15].

2.2.11 Statistics

Data analytics in football has grown exponentially in the last decade. Clubs are using many metrics to measure a player's qualities. As shown in conditions that affect performance, different metrics must be used for players in different positions. Basic metrics include short conversions, passes, dribbles, tackles, goals, assists, and clean sheets. However, a more in-depth analysis can consider the context of these metrics, such as progressive dribbles, carries into the opponent's box, and passes into the final third.

New advanced statistics have appeared in the world of football, such as expected goals (xG), expected goals on target (xGOT), expected Assists (xA), and expected pass completion (xP) [17]. All of these statistics are created with a probabilistic model developed for the purpose of providing further insight into underlying metrics. In the case of xG, the idea behind the model is ”given a set of shot characteristics of predictors; the model estimates a probability for the observed shot" [18]. These stats are convenient because they can be trained on data from the top leagues in Europe. This raises the standard of the statistics and provides a more insightful view of how a player is performing. Many top players will consistently outperform these advanced statistics because they are better in that area than the players around them.

2.2.12 Data

Data selection is the foundation upon which data-driven analyses, machine-learning models, and strategic decisions are built. The data chosen for analysis directly influences the results and insights derived from the analysis. Irrelevant or poor-quality data can lead to poor conclusions and misguided actions. Selecting the correct data makes for efficient use of resources, including time, computational power, and human effort. Focusing on relevant data streamlines the process and keeps it efficient.

The process of selecting data has its challenges and complexities. Acknowledging these challenges is vital for choosing the most accurate data. In the era of big data, the sheer volume of available data can be overwhelming. Deciding which portions of this vast data landscape are relevant can be daunting. Data quality is a persistent concern. Data may contain errors, inconsistencies, or missing values that must be addressed during selection. Data can be sourced from various places, including databases, Application Programming Interfaces (APIs), external vendors, and internal systems. Choosing the most suitable data sources requires careful consideration of accessibility, reliability, and cost.

Selecting features necessary for the project is crucial in machine learning as it objectively chooses the most informative attributes from a dataset.

2.2.13 Challenges

There are some challenges when evaluating and measuring a player and their abilities. Taking a data-focused approach can lead to better outcomes. However, challenges are still present, such as a need for more data, context, and experience to interpret the data; football is a challenging game to collect data on, adapt to new statistics, and split the individual performance from the team. The problem with advanced metrics is that the team around them can heavily influence individuals; strikers in better teams will get more chances and have better xG and more goals than in lower-class teams [19]. There is a reluctance to move towards an entirely data-driven approach. Many people within football have a traditionalist view of the game and appreciate and are comfortable with a more human approach [19]. Data quantity and quality need to be increased in areas outside of top leagues and developed countries. In order to take a data-driven approach to scouting across the world, infrastructure and technology usage need to increase.

Chapter 3

Problem - Player evaluation using machine learning to assess the ability of a player

3.1 Problem Definition

Player evaluation in sports, particularly football, has significantly transformed. Historically, the process relied heavily on in-person scouting and intuitive judgment. Experienced scouts, using their keen observation and expertise, played a crucial role in identifying and assessing talent. However, this approach had inherent limitations, including susceptibility to biases, reliance on subjective judgment, and the variable quality of observation, such as a player having a good or bad game.

As football grew into a global phenomenon with substantial financial stakes, the need for a more robust, objective, and data-driven approach to player evaluation became clear. The evolution of scouting has mirrored the evolution of the game itself – moving from a largely intuitive process to one underpinned by statistical analysis and data-driven insights.

Currently, the landscape of player evaluation is predominantly shaped by statistical analysis. Data-driven approaches have provided a more objective framework for assessing player abilities. However, this reliance on statistics is challenging. The correct interpretation of statistical data requires expertise; even then, it can be prone to misinterpretation. Moreover, while providing valuable insights, statistics only sometimes capture the nuanced and dynamic nature of a player's abilities and potential.

This project aims to address these challenges by leveraging machine learning to develop an algorithm more comprehensively inferring a player's abilities and qualities. The goal is to create a tool beyond traditional statistical analysis, offering a more nuanced and holistic assessment of a player's level.

3.2 Objectives

3.2.1 Data collection

This phase is crucial as it lays the groundwork for the entire project. High-quality, extensive data collection is imperative, focusing on a variety of relevant features and indicative of player performance. The challenge here lies in ensuring the data's quality, relevance, and comprehensiveness. It is not just about quantity but also about the variety and depth of data, covering various aspects of performance. This step may involve collating data from multiple sources, ensuring a diverse range of data points that can contribute to a more robust analysis.

3.2.2 Feature selection

Once data is collected, the focus shifts to feature selection. This step is critical in determining the effectiveness of the algorithm. It involves sifting through the collected data to identify features that strongly correlate with performance metrics. This process is not arbitrary; it requires a foundation in extensive research and literature review. The aim is to determine which features are most relevant and how they should be weighted in the algorithm. This phase ensures that the algorithm focuses on the most impactful data, enhancing its accuracy and relevance.

3.2.3 Develop algorithm

The development of the algorithm is at the heart of this project. This phase involves designing a sophisticated algorithm capable of effectively analyzing the collected data to provide accurate evaluations of player performance. The design must focus on efficiency, accuracy, and user-friendliness. It might involve integrating advanced machine learning techniques and statistical methods to enhance the algorithm's predictive accuracy. The challenge here is to create an algorithm that is accurate in its current form and scalable and adaptable for future enhancements.

3.2.4 Accuracy testing

Testing the algorithm's accuracy is a critical phase to ensure its reliability. This involves using historical data as a benchmark to validate the algorithm's evaluations. The testing should be comprehensive, covering different player positions and roles, as performance metrics can vary significantly based on these factors. This phase helps fine-tune the algorithm, identify any biases or inaccuracies, and make necessary adjustments to improve its performance.

3.2.5 Versatility testing

The versatility of the algorithm is tested by applying it across different contexts. This includes testing across various leagues, including women's football, and potentially extending it to other sports. The aim is to ensure the algorithm is manageable and can handle diverse datasets and performance indicators. This phase assesses the algorithm's flexibility and adaptability, ensuring it can provide accurate evaluations in various scenarios.

3.2.6 Discuss future research

Finally, discussing potential future research based on the project's results is crucial. This involves analyzing the outcomes, identifying limitations, and exploring areas for improvement. The discussion should also focus on how the algorithm can evolve with further research, technological advancements, and changes in sports analytics. This phase sets the direction for future work and advancements in the field, highlighting new features, methodologies, or areas that could be explored to further enhance the algorithm's capabilities.

3.3 Functional Requirements

The core objective of this project is to develop a machine-learning algorithm for player evaluation in sports. To achieve this, a comprehensive data collection process must be implemented. This involves gathering substantial, high-quality data from multiple leagues, including women's football and other sports, rich in various performance indicators. The data should be diverse and encompass various features relevant to player performance.

Once data is collected, the next step is feature selection. This process involves identifying key performance indicators (KPIs) most indicative of a player's abilities. These KPIs should be selected based on extensive research and statistical analysis, ensuring they genuinely reflect player performance. The significance of each feature must be evaluated, and appropriate weights assigned to them must accurately reflect their importance in the overall performance assessment.

The development of the machine learning algorithm is central to this project. The algorithm must analyze the collected data comprehensively and provide an accurate evaluation of a player's abilities. This algorithm must be adaptable and capable of assessing players across different positions, sports, and leagues, ensuring broad applicability and inclusivity.

Testing the algorithm's accuracy is crucial. This can be achieved by using historical data as a benchmark. The accuracy testing must consider different positions, recognizing that the relevance of features may vary across sports roles. Additionally, the versatility of the algorithm should be tested across various leagues and sports, including both men's and women's sports, to ensure its broad applicability.

Finally, the project should include a component that discusses future research avenues. Based on the outcomes, potential areas for further development and enhancement of the algorithm should be identified, focusing on improving its accuracy, versatility, and adaptability to the evolving dynamics of sports.

1. Data collection
    - Ability to process and clean data.
2. Feature selection
    - Ability to choose different features.
    - Ability to alter weighting for features.
3. Develop Algorithm
    - Design algorithm capable of analyzing the data and evaluating a players performance.
    - Functionality to generate a report on the evaluation of the player.
4. Accuracy testing
    - Ability to test the accuracy of the algorithm against historical data.
    - Ability to test the accuracy for different player positions.

5. Versatility testing
    - Ability to test across different leagues.
    - Ability to test across gendered sports.
    - Ability to test across different sports.

3.4 Non-Functional Requirements

In terms of non-functional requirements, the project must prioritize accessibility and usability. The final report and any software tools developed should be easily understandable and user-friendly, catering to sports professionals and non-experts. This includes a clean and intuitive interface for any software application.

Originality and innovation are key. The project should build upon existing research and methodologies but must ensure that its outcomes are distinct and contribute new insights or improvements in player evaluation. The project's findings must not be merely a repetition of existing works but rather an advancement of the field.

Data security and privacy are paramount. The project must implement stringent data security measures to protect sensitive information and comply with all relevant privacy laws and ethical standards. This includes safeguarding personal data and ensuring confidentiality.

Scalability and maintenance of the system are essential considerations. The design should allow for easy updates and expansion as new data sources and sports analytics methods emerge. A maintenance plan for the algorithm and any associated software should be established, including provisions for regular updates and bug fixes.

Finally, comprehensive documentation and reporting are vital. The project should maintain detailed records of the research process, algorithm development, and testing methodologies. The final report should clearly articulate the project's findings, methodologies, and implications, presented in a structured and understandable manner.

3.5 Non-Functional Requirements

1. Performance
    - The system should be able to process a large quantity of data in a fast and efficient manor.

2. Scalability

   - The system should be able to scale to include a greater amount of data and features.

3. Compatibility

   - The system should be capable of dealing with various formats for data and can continue to operate.

4. Usability

   - The system should not be difficult to operate.

Chapter 4

Implementation Approach

4.1 Risk Assessment

- Consequence

    1. Minor: Results have a negligible or trivial impact on the project or system, with no significant long-term effects, and are easy to manage or rectify.
    2. Major: This leads to substantial disruption, requiring significant resources or time to manage and recover, with a lasting impact on project outcomes.
    3. Critical: Causes severe and extensive damage to the project or system, potentially resulting in long-term impairment and requiring extensive time to fix.
    4. Fatal: Results in catastrophic outcomes, leading to complete project failure or irreversible damage, often beyond the scope of recovery.

- Frequency

    1. Rare: The event is unlikely to occur during the lifetime of the project or system.
    2. Remote: The event could occur but only under unusual circumstances.
    3. Occasional: The event has a moderate chance of occurring, likely to be encountered at some point.
    4. Probable: The event is likely to occur several times during the lifetime of the project or system.
    5. Frequent: The event is expected to occur frequently, a common occurrence in the project or system lifecycle.

4.1.1 Data-Related Risks

- Consequence: 3
- Frequency: 4

The primary risk is data accessibility. Securing extensive, high-quality datasets from reliable sources can be challenging. Sports data, exceptionally detailed player performance data, might be proprietary or expensive to acquire. Even if accessible, these datasets can be vast and complex, demanding significant preprocessing and cleaning to be usable, which can be a daunting task for someone with limited experience. Another primary concern is data bias. In sports analytics, data might inherently contain biases based on historical preferences, regional play styles, or under representation of certain groups (like women or players from less popular leagues). These biases could skew the algorithm's predictions if unaddressed, leading to inaccurate player evaluations. Lastly, there is the risk of misinterpreting data due to lack of experience. Accurately understanding and interpreting sports performance data requires technical skills and a deep understanding of the sport itself. Misinterpretation could lead to selecting inappropriate features for the algorithm or incorrectly weighing these features, impacting the outcomes.

4.1.2 Technical Risks

- Consequence: 1
- Frequency: 5

Addressing technical risks is incredibly important, as these risks can directly impact the feasibility and success of the project. One of the primary technical risks involves the development of the machine learning algorithm itself. There are many challenges involved when building complex, high-performance algorithms. These challenges are serious risks that must be considered. These challenges can lead to problems in creating an accurate and efficient algorithm. Key issues include properly selecting and tuning the machine learning model, handling over fitting or under fitting, and ensuring the algorithm can generalize well to new, unseen data. Another significant risk is related to data processing and feature selection. The project demands a sophisticated understanding of how to preprocess data, handle missing or noisy data, and select features most indicative of a player's performance. Incorrect feature selection or poor data preprocessing can lead to sub-optimal algorithm performance. Furthermore, scalability and performance optimization are crucial concerns. The algorithm needs to process potentially large datasets efficiently. Access to high-powered computing resources might be limited, and developing an algorithm that operates efficiently on available resources can take time and effort. This limitation can become even more pronounced as the project scales or the algorithm becomes more complex.

4.1.3 Project Management Risks

- Consequence: 2
- Frequency: 3

The foremost risk is the mismanagement of time, and dedicating sufficient time to a project of this scale can be challenging. There is a risk of underestimating the time required for different aspects of the project or becoming sidetracked, such as data collection, algorithm development, and testing. This can lead to rushed work or missed deadlines, impacting the quality and success of the project. Another significant risk is scope creep, where the project's requirements gradually increase beyond the initial plan. This might occur due to the evolving nature of the project, new ideas, or the need for clear initial objectives. Scope creep can lead to an unmanageable workload and divert resources and attention from the project's core objectives. Furthermore limited experience in managing complex projects, leading to planning, resource allocation, and risk management issues can result in inefficiencies, conflicts, and delays.

4.2 Methodology

4.2.1 Data Collection and Preparation

The data collection phase is foundational to the project. It involves identifying reliable sources for player performance data, including sports analytics databases, public datasets, and league statistics. Once the data is gathered, it undergoes thorough cleaning and preprocessing. This step ensures data quality and usability, involving tasks like handling missing values, correcting errors, and standardizing data formats. Data exploration is also conducted to gain initial insights into the datasets, such as identifying patterns, outliers, and basic statistical properties, which aids in understanding the data's structure and potential challenges it may pose.

4.2.2 Feature Selection and Engineering

In this stage, the focus shifts to isolating the most relevant features for player evaluation. This begins with identifying Key Performance Indicators (KPIs) that accurately reflect player abilities in the chosen sport. Then, feature engineering techniques are employed to create new features that could provide additional insights, such as aggregating statistics over a season. The selection of features is a critical step involving statistical techniques to determine the most impactful variables. This process is iterative and requires a balance between including informative features and avoiding overly complex or irrelevant ones.

4.2.3 Model Development

The model development phase is at the heart of the project. Selecting the right machine learning algorithms is based on the nature of the data and the project's specific goals. Common choices include regression models, decision trees, or neural networks. The model is trained using a portion of the data, carefully adjusting parameters to optimize the performance. Cross-validation techniques are employed to evaluate the model's effectiveness, helping to prevent issues like over fitting and ensuring that the model can effectively and fairly deal with new data.

4.2.4 Testing and Validation

Testing the model against historical data is critical to validate its accuracy and reliability. This involves applying the model to past player performance data to see how well it predicts their outcomes. Additionally, applying the model to real-world scenarios helps assess its practical applicability and versatility across different contexts within the sport. This step is essential to demonstrate the model's utility for potential real-world applications.

4.2.5 Documentation and Reporting

Maintaining clear documentation throughout the project is vital. This documentation includes detailed records of the methodologies, data handling processes, model development steps, and evaluation criteria. The final report compiles all these elements, presenting the project's findings, methodologies, results and insights in a structured and accessible way. It also discusses the project's limitations and areas for improvement, providing a clear overview of the project's outcomes and contributions.

4.3 Implementation Plan Schedule

- Data Collection and Preparation (Week 1-3)

  Sourcing and collecting data

  Data cleaning and prepossessing

- Feature Selection and Beginning Algorithm Development (Week 5-6)

  Selection of Key Performance Indicators (KPIs)

  Developing first edition of the algorithm

  Basic testing

- Algorithm Enhancement and Testing (Week 6-8)

  Continuous development of algorithm

  Testing performance

  Testing accuracy

  Testing bias

- Versatility testing (Week 9)

  Source alternative datasets

  Using different datasets

  Testing across different sports

  Testing across different age groups

  Testing across different genders

- Documentation. (Week 9-12)

  Documentation

4.4 Evaluation

In order to create a plan that can accurately evaluate the project, the functional and non-functional requirements must be the plan's focus. Each requirement must be viewed as a judgment on whether it adequately satisfies the requirements.

1. Data collection

   - Evaluation Criteria:

     (a) Data Relevance

(b) Data Quality

(c) Data Volume

- Method: For data collection, we will implement automated quality checks and error rate assessments to ensure the relevance and quality of the data. Additionally, the volume of data will be compared against established standards within similar projects to gauge adequacy.

2. Feature selection

   - Evaluation Criteria:

     (a) Statistical Significance

     (b) Feature Weighting

     (c) Usability

   - Method: In feature selection, statistical tests like ANOVA and chi-square will be utilized to determine the statistical significance of each feature. Weight analysis algorithms will be applied to assess the impact of features, and their usability will be evaluated based on internal testing and alignment with project goals.

3. Develop Algorithm

   - Evaluation Criteria:

     (a) Provides Result

   - Method: The algorithm development phase will involve rigorous internal testing in controlled environments, ensuring the algorithm delivers accurate and consistent results. Various inputs will be iteratively tested to assess the algorithm's robustness and reliability.

4. Accuracy testing

   - Evaluation Criteria:

     (a) Historical

     (b) Cross-Validation

     (c) Bias

   - Method: For accuracy testing, the algorithm's outputs will be compared against historical data to determine accuracy. K-fold cross-validation methods will be used to evaluate performance across different data subsets. Furthermore, internal tools will be utilized to analyze the outputs for potential biases.

5. Versatility testing

- Evaluation Criteria:
    (a) Different Data Sets
    (b) Different Sports
    (c) Different Genders
    (d) Different Age Groups
- Method: Versatility testing will include applying the algorithm to various datasets to check its adaptability across different sports, genders, and age groups. This will help assess the algorithm's flexibility and applicability in diverse scenarios.

6. Performance
    - Evaluation Criteria:
        (a) Processing Speed
        (b) Reliability
    - Method: The algorithm's performance will be evaluated by measuring processing speeds across datasets of varying sizes and conducting stress tests under heavy loads. The algorithm's efficiency will be analyzed based on resource utilization, including CPU and memory.
7. Scalability
    - Evaluation Criteria:
        (a) Data Volume
        (b) Ability To Improve
    - Method: Finally, in assessing scalability, we will test the algorithm's capability to handle increasing volumes of data and its adaptability to include improvements and expansions. This will be achieved through incremental data volume testing and evaluating the algorithm's modular design for easy integration with new features or datasets.

Chapter 5

Implementation

5.1 Sprint Review

5.1.1 Sprint 1 (3 weeks)

5.1.1.1 Plan

The goal for this sprint is to create the first iteration of the model.

In order to do this, there are three things that must be done. Acquire data, clean and process the data so that it can be used and create the first iteration of the model.

When searching for data several criteria needed to be met, this included the following.

1. Quantity

   It should provide enough data to train the model and minimize bias.

2. Variety

   It should be gathered from various leagues.

3. Quality

   It should provide many different metrics on ability.

4. Usability

   It should be usable or easy to clean.

5. Quantifiable

   All metrics should use numbers.

6. Timely

   It should be of recent history.

The reason for this is to gather a dataset that provides a fair representation of footballers with a particular focus on England, Spain, Italy, Germany, and France, which are known as the Big Five as they ”dominate world football with revenues and resources greater than any other collection of leagues” [20].

Processing and cleaning the data is a necessary step in order to use the data we acquire. We need to do this so we can check for discrepancies and fix any formatting issues that may cause errors.

Finally, we plan to make a Regression model that will utilize multiple metrics to give a prediction on a target variable. The model’s accuracy and which algorithm chosen is not a concern in this Sprint. However, the model should still produce a valid result, meaning an output that can be used to predict the target variable.

5.1.1.2 Achieved

Football Manager was chosen as the source of data. It has been credited by many within football for its accuracy and even signed a deal with Everton in 2008 to allow the club to use its database to search for players and staff [21]. Furthermore, Football Manager provides a straightforward system for quantifying metrics on a scale of 0-20. Football Manager covers all European leagues and more internationally. It is also updated every year, and all data is presented neatly. This was the best dataset we could find, as it fulfilled my criteria, was of higher quality, and was more consistent than any other source. There was no cost to acquiring the data because we already possessed the game.

We used Google Sheets to process and clean the data. The data set was initially acquired in a comma-separated values (CSV) format, and this was the best and easiest way to display and clean the data. During the cleaning process, we reformatted values that used characters or words to represent amounts and replaced them with their digit counterpart, for example, ”100k” to ”100,000”. Heights are represented as centimeters (cm), and most footballers are under 200cm, which is easily scaled to 0-20 and is used by all other metrics; we decided to represent any football taller than 200cm as 200cm. I did this because it makes it consistent with other data; there are very few footballers taller than 200cm, and we do not think it will have a notable impact.

The algorithm we choose to begin development with is Ordinary Least Squares (OLS), which is a ”common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression)" [22]. It is also a "widely used tool in econometrics. It allows estimating the relation between a dependent variable and a set of explanatory variables" [23]. The target variable for the model is the transfer value; given this, we believe this algorithm is acceptable and fulfills my requirements. Using this model and dataset, we achieved a predictive value for the transfer value.

1. First iteration - 32.99% R-squared
2. First iteration - 10,802,188.19 RMSE
3. First iteration - 6,593,061.02 MAE

Using several features chosen based off intuition such as Height, Age, Technical and Stamina.

5.1.1.3 Evaluation

For this sprint, we had three goals. Acquire data, clean and process the data, and create a model using this data. During this sprint, we reviewed several different datasets and successfully sourced one that fulfilled all my criteria. We were able to process and clean the data so that it can be used as needed. Finally, we were able to create a model that was able to use this data to predict the target variable. Given that we achieved the goals for the sprint, we think this was successful.

5.1.2 Sprint 2 (3 weeks)

5.1.2.1 Plan

The goal for this sprint is to refine the current model.

In order to refine the current model, there are two propositions. These are to split the model into three model for the positions' defence, midfield and attack and to refine the features used in order to provide the set of features with the best predictive capability.

The reason for splitting the model into three is that each position has unique features relevant to it. By splitting the model, we can provide a custom feature set for each position that will allow us to target those players more effectively and provide better predictions.

5.1.2.2 Achieved

To further refine the model, it was decided to split it into three different models: defense, midfield, and attack. This decision was based on the literature's findings that different traits for positions indicate their value. By splitting the model, a more granular approach could be taken when choosing features to provide the best prediction. For example, finishing features related to shooting are much more indicative of an attacker's value than a defender's.

The models proved to be more effective when split and using their own set of features than when using one model. This can be seen here where a set of features oriented towards strikers was used on the models. The striker oriented model provided the best result.

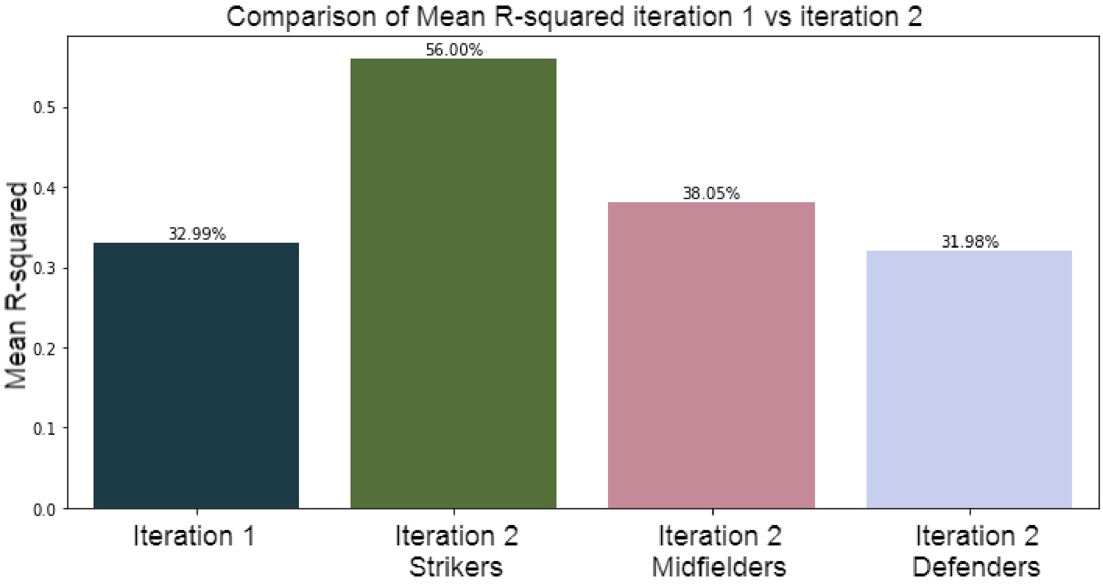


FIGURE 5.1: Accuracy Iteration 1 vs Iteration 2

1. First iteration - 32.99% R-squared
2. Second Striker iteration - 56.00% R-squared
3. Second Midfielder iteration - 38.05% R-squared
4. Second Defender iteration - 31.98% R-squared

However, we felt there was potential for further improvement during this sprint. During the implementation phase, the target variable for the project changed from predicting the ability to transfer value because of limitations. Due to this, we reviewed our methodology for predicting the target variable. Further investigation was carried out, and,we were not satisfied with our methodology upon review. The decision was made to revert to a singular model, now using the positions as a feature. Furthermore, although satisfied with the research and the fulfillment of criteria, it was deemed necessary to test multiple algorithms to prove that the most fitting algorithm was chosen [24].

During this sprint, several challenges presented themselves. Time management became an issue, and not enough time was put into development during the sprint. This led to a lack of progression in the project's development. Poor foresight and planning caused a necessary revision in the project's methodology. This was a setback for the project and a waste of time.

5.1.2.3 Evaluation

The goal for this sprint was to refine the model. Although steps were taken towards refining the model, we felt that it was not the best path forward, and revisions were made. This means that the sprint could not meet its goal of refining the current model. However, a new approach is being taken in light of this and is expected to provide better results.

5.1.3 Sprint 3 (3 weeks)

5.1.3.1 Plan

The goal for this sprint is to identify the best model that can achieved with our current research.

In light of discoveries during the second sprint, this sprint intends to thoroughly exhaust all avenues to build open the research and development already done to identify the best method for predicting the player's transfer values, including the algorithm and features. Furthermore, the decision was taken to incorporate positions as a feature.

5.1.3.2 Achieved

For this sprint, we wanted to ensure that the path we were traversing most effectively affected our goals. In order to do this, we decided to evaluate our model using the OLS algorithm against several other algorithms to demonstrate it was the best choice. To do this, OLS was chosen as the baseline against which to evaluate the other models. The metric used to evaluate the accuracy was mean R - squared, the ”statistical measure representing the proportion of the variance for a dependent variable” [25]. The models we chose to evaluate were gradient-boosted regressor (GBR), Elastic Net (EN), K Nearest Neighbours (KNN), Random Forest Regression (RFR), and Decision Tree Regression (DTR).

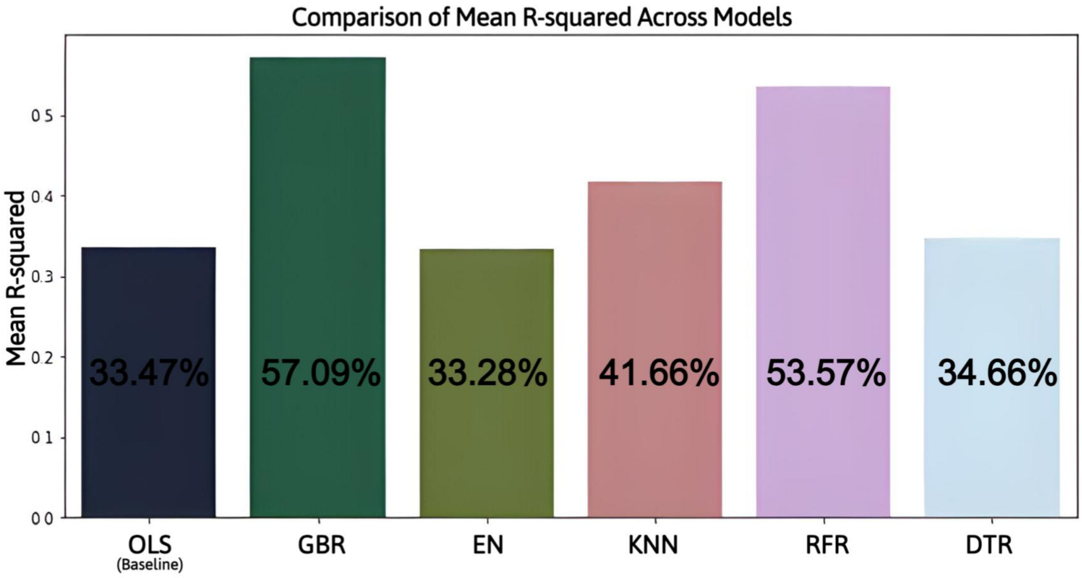


Figure 5.2: Accuracy Of Models

From these models, we found that GBR was the most accurate, with a mean R-squared of 57.09% more significant than any other and a 23.62% increase on the baseline model of OLS. We refined the model with these findings to cater to this algorithm and maximize our results. There was significant room for improvement, and a mean R-squared of 81% was achieved. We are satisfied with this accuracy. This was done by limiting the leagues to the five largest in Europe.

1. English Premier League
2. German Bundesliga
3. Spanish La Liga
4. French Ligue 1
5. Italian Serie A

We also found that limiting the age range negatively affected the accuracy and that incorporating the positions as a feature provided a more consistent result.

5.1.3.3 Evaluation

This sprint aimed to identify the best algorithm for the model and achieve an acceptable accuracy for the model. We successfully achieved this goal by evaluating multiple algorithms and refining our model through rigorous trial and error processes and research into similar research to review their methodology. This sprint has completed the algorithm's foundation and set a solid platform for continuous development.

5.1.4 Sprint 4 (3 weeks)

5.1.4.1 Plan

The goal for this sprint is to improve our model and create a way to interact with it.

For this sprint we want to continue development on the model so that it can produce an acceptable accuracy score. Secondly we want to create something that can provide an interactive experience with the model, this will likely to shape in the form of a webpage or application.

5.1.4.2 Achieved

During this sprint, we were able to refine the model. It can now explain 80% of the variability of the transfer value for a player and has an MAE of 2,253,920.87 and RMSE of 4,414,550.74. We are satisfied that this iteration of the model is satisfactory and can claim to predict the transfer value of a player. Furthermore, we created an interactive webpage where the model can be observed. On this webpage, features and leagues can be chosen to generate and provide different perspectives on the model and its relationship with features. The webpage shows the importance of features in attaining its predictions. Two additional leagues were added to the webpage: the Dutch Eredivisie and the Portuguese Primera Liga.

5.1.4.3 Evaluation

This sprint aimed to improve the model and provide a method to interact with the model. We produced a model during this sprint that now provides a satisfactory accuracy score. Creating the interaction webpage offers an enjoyable way to interact with the webpage. This sprint was successful, and we achieved all of our goals.

5.2 Actual Solution Approach

This section explores the methods and techniques used to develop and refine our models for predicting player performance. This part of our project involves several steps where we carefully choose, test, and improve our approaches based on the data we have collected and prepared. Through an iterative process of trial and error, guided by our project reviews, we adapt and refine our methodology to ensure that our model is both accurate and effective.

5.2.1 Data Collection and Preparation

This section focuses on gathering and preparing the data necessary to build good models that can predict player performance. This step is significant because it ensures that we have reliable and detailed data with which to work. We collect data from various trusted sources and ensure it is clean and ready for our models.

5.2.1.1 Data Collection

For the data collection, several criteria needed to be met for the dataset to fulfill our needs. These criteria included quantity to ensure we had enough data, variety so we were not focusing on one area and had a fair representation, quality to make sure that the data we were using was consistent and reliable, usability so that data was in a usable format and was presented in an organized manner, quantifiable so we could use the data in the project while making as little inferences about the data as possible and finally timely which means that data needed to be recent because to due to the nature of transfer valuation and the market there is a relationship with time.

Selecting the data was a process undertaken in the project's implementation phase; however, upon development and further investigation into the data source that we had planned to use. We discovered it fulfilled only some of these criteria, and thus, we had to continue searching. Through the suggestion of our supervisor, we investigated Football Manager, a game developed by Sports Interactive and published by SEGA. The database in this game is reputable, curated, and created through real football scouts and has been used professionally by football clubs, as referenced in Section 5.1.1.1 during the Sprint Review.

The dataset contained 25,087 football players from many different leagues across the world. It also contained 56 features for each player, which included their name, nationality, position, club, division, value, age, height, speed, stamina, strength, work ethic, passing, tackling, shooting, and many more related to niche parts of the game. Although we were only interested in the top five leagues in Europe, we had 5,420 players. The range for the target variable in this data was 91,000,000 to 350. We included two more leagues in the interactive web version, which has 5,731 players and 46 features that can be used.

5.2.1.2 Data Preparation

The dataset we collected, integral for predicting football players' transfer values, necessitated significant preprocessing to make it suitable for use in our predictive models. The main tasks involved formatting critical columns such as 'transfer value' and 'height', which were initially inconsistent across the dataset. Google Sheets was selected as the tool for this phase of data preparation due to its functionality and ease of use.

5.2.2 Feature Selection and Engineering

In this section, we discuss the specific steps we took to develop our predictive models. We'll outline how we selected the appropriate modeling techniques, set up our experiments, and adjusted our methods based on the outcomes. This part of our project is crucial as it shows the practical application of our theories and the adjustments made to refine our models, ensuring they are both accurate and helpful in predicting player performance.

5.2.2.1 Feature Selection

We included every feature from the dataset. While some were more important than others, the added context of these features benefited the model. There were some concerns over particular features, such as age and position. These concerns were due to outliers and the unique nature of the position feature. The project is targeted towards footballers and not a specific position, so in this sense, we thought it best to use it as a feature and not a reason to split the model to target positions.

5.2.2.2 Feature Engineering

In order to manipulate the data, we used the libraries pandas and numpy. These libraries are commonly used when working with data analytics and were of great use and convenience. When working with the feature age, we considered applying constraints;

in order to do this, we filtered out any players below the age of eighteen and above the age of forty. We attempted, through trial and error, to test for a better combination of ages, but we did not find anything meaningful. Upon removing this constraint, we found that the model performed better. Similarly, for a footballer's position, we attempted to apply constraints to work with a specific position to provide customized features for that possession, expecting it to bear greater results. This did not work how we anticipated, and we found greater results when utilizing position as a feature. We did end up applying weighting to the positions to reflect the nature of more attacking players being worth more than defensive players. We did this by viewing the positions a player can play and assigning the corresponding value to the most attacking position they could play. Through trial and error, this has had a significant positive impact.

5.2.3 Model Development

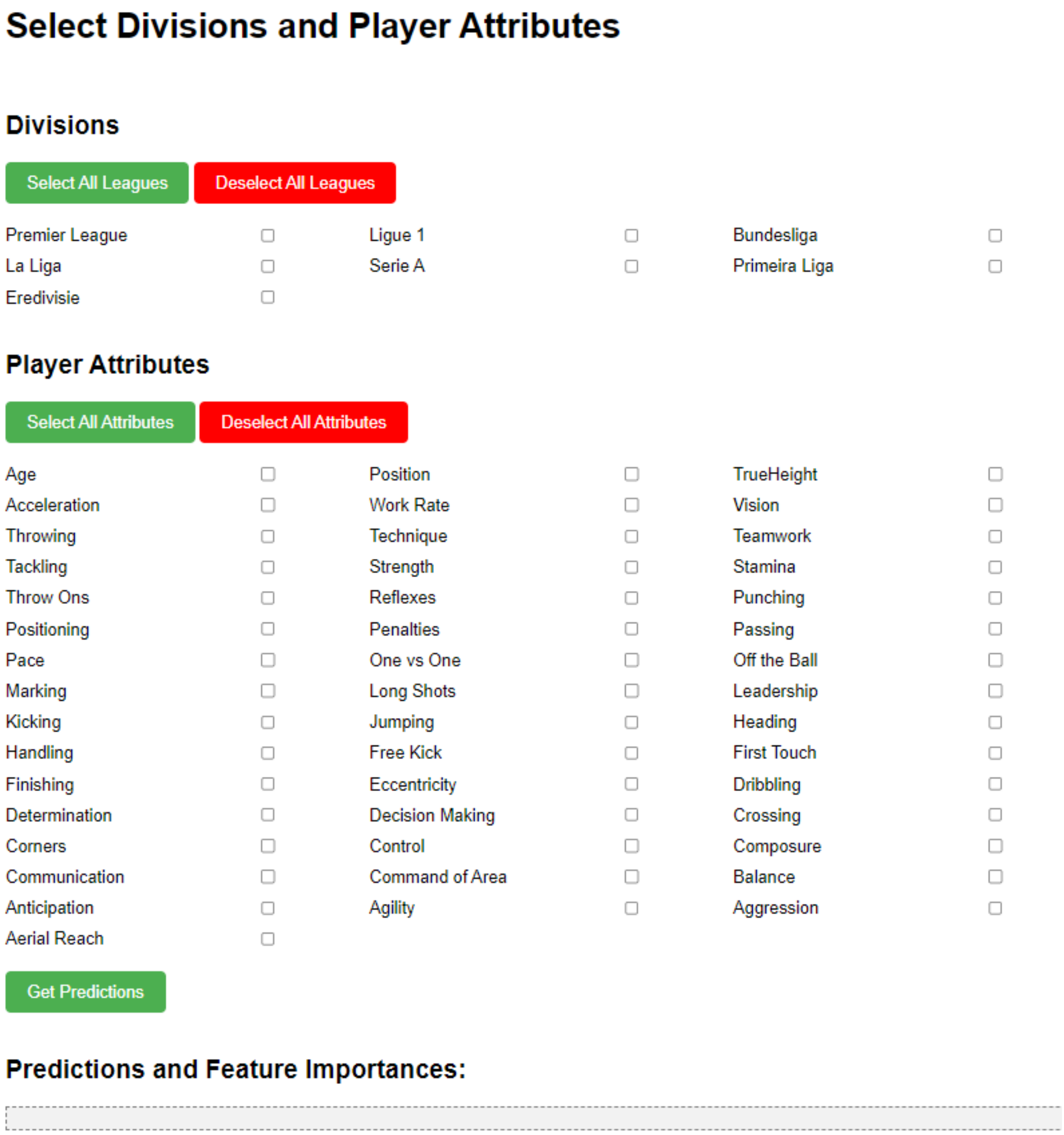
In this section, we describe the development of our models to predict player performance. We explore the techniques and approaches used to build and refine these models, focusing on algorithm selection, feature engineering, and integrating diverse data sources. This part of the project is essential as it outlines the foundational work that enables our models to forecast outcomes accurately based on the data.

5.2.3.1 Model

The model was developed using Python, a versatile programming language favored for data analytics. Specifically, we utilized the scikit-learn (sklearn) library, an open-source and widely acclaimed library for machine learning, to implement and test our machine learning models. We employed several key modules within this library to enhance our model's functionality and reliability. We utilized the KFold module from scikit-learn to perform cross-validation, ensuring our model was robust and generalizable across unseen data. This validation method divides the dataset into k consecutive folds, allowing us to iteratively train and test the model across different data segments and thus prevent overfitting. Additionally, we used the GradientBoostingRegressor module to implement the regression model. This approach not only facilitated the development of a robust predictive model but also ensured that our findings were grounded and based on proven and trusted statistical methods, which reassures the credibility of the findings.

5.2.3.2 Display

Using the Flask web framework, we developed a simple yet interactive webpage that serves as a user interface for our predictive model. This webpage allows users to actively engage with the model by selecting various player features, such as age, position, and technical skills. Through this interface, users can observe how different inputs affect the predicted transfer value, providing a hands-on experience of the model's functionality in real time. To enrich the user experience and extend the scope of our research and development, we included data from two additional football divisions, these are the Dutch Eredivisie and the Portuguese Primeira Liga. Including these leagues broadens the dataset and provides greater insights. This expansion allows users to explore various scenarios and investigate the relationships. Furthermore, the interactive webpage is designed to be user-friendly, catering to experts and novices. It incorporates intuitive controls for feature selection and clear visualizations of the model's outputs. This approach demonstrates the model's capabilities. By integrating these features into the webpage, we aim to facilitate ongoing improvements and encourage further academic and practical exploration of predictive modeling in sports analytics. This platform has the ability to continuously evolve as we integrate more data and enhance the model.



Select Divisions and Player Attributes

Divisions

Select All Leagues | Deselect All Leagues

Premier League	□	Ligue 1	□	Bundesliga	□
La Liga	□	Serie A	□	Primeira Liga	□
Eredivisie	□				

Player Attributes

Select All Attributes | Deselect All Attributes

Age	□	Position	□	TrueHeight	□
Acceleration	□	Work Rate	□	Vision	□
Throwing	□	Technique	□	Teamwork	□
Tackling	□	Strength	□	Stamina	□
Throw Ons	□	Reflexes	□	Punching	□
Positioning	□	Penalties	□	Passing	□
Pace	□	One vs One	□	Off the Ball	□
Marking	□	Long Shots	□	Leadership	□
Kicking	□	Jumping	□	Heading	□
Handling	□	Free Kick	□	First Touch	□
Finishing	□	Eccentricity	□	Dribbling	□
Determination	□	Decision Making	□	Crossing	□
Corners	□	Control	□	Composure	□
Communication	□	Command of Area	□	Balance	□
Anticipation	□	Agility	□	Aggression	□
Aerial Reach	□				

Get Predictions

Predictions and Feature Importances:

FIGURE 5.3: Interactive Model Webpage

5.2.4 Testing

In this section, we will test and validate our predictive model to ensure their effectiveness and reliability. This crucial phase involves applying the models to various datasets to evaluate their accuracy and applicability. We will discuss the different techniques used for testing, the metrics employed to evaluate model performance, and the subsequent adjustments made based on these evaluations.

5.2.4.1

Testing The project utilized three key performance metrics for testing: RMSE, MAE, and R-squared. These metrics were applied to every model iteration, providing a clear and quantifiable indication of its accuracy throughout development. We could track the model's progress by evaluating these metrics at each iteration and make data-driven decisions to enhance its predictive accuracy. To ensure that the model was robust and generalized well to unseen data, we employed k-fold cross-validation with a specific focus on using an 80:20 split for training and testing the model. This method was consistently used across all iterations, allowing each data segment to serve as training and testing data at different points. This rigorous approach not only helped validate the model's effectiveness across various subsets of data but also minimized the risk of overfitting, thereby enhancing the reliability of our findings.

5.3 Difficulties Encountered

Throughout the project we encountered a variety of difficulties , which ranged from easy to hard to overcome. Earlier in the project we identified three key areas in which we foresaw risks appearing. These were:

- Data-Related Risks
- Technical Risks
- Project Management Risks

These difficulties varied widely in their nature and complexity, impacting our development process to differing degrees. In order to clearly articulate the spectrum of issues faced and how they were addressed, we have categorized these challenges into three distinct levels:

- **Easy**: The problem did not take much time to overcome and the original approach and outcome were achieved.
- **Medium**: The problem took some time to overcome and the original approach may have changed but the outcome was achieved.
- **Hard**: The problem posed such an issue that it was not able to be overcome and the functionality or the outcome envisioned was not achieved or was changed greatly.

Each of these categories reflect the effort required and the success obtained from overcoming these difficulties, which ranged from simple solutions to more complex and difficult approaches.

5.3.1 Acquiring a dataset

(Hard, Data Related)
For the project, acquiring a dataset that fit our needs proved to be the most difficult task, as several criteria needed to be met to satisfy our needs. We needed a dataset representing the entire spectrum of features a footballer could possess. It needed to be significant, it needed to have real-world applicability, it needed to be of a high standard and consistent in terms of its rating, and in terms of all data being available, it needed to be diverse, and the features needed to be quantifiable. These needs pose a considerable challenge and severely limit the options available.

5.3.2 Data Preprocessing

(Easy, Data Related)
Data preprocessing was an issue we faced with all available datasets. No data was readily available to be used in a regression model, which meant doing some preprocessing before manipulating the data in python. Some elements of that data needed to be formatted before use; this included transfer value and height. In order to mitigate this, we manipulated the data in Google Sheets, which is an alternative to Microsoft excel. We used formula-based data manipulation to remove any unnecessary characters or to format values in a consistent format.

5.3.3 Time Management

(Easy, Project Management Related)
Being able to work on the project within an agile context consistently proved challenging throughout development. Due to many factors external to the project's demands, such as course work, personal commitments, and unforeseen circumstances, time management was a significant challenge. To address this, we implemented a flexible but structured schedule that allowed for adjustments while maintaining progress toward milestones. We adopted agile methodologies, specifically Scrum practices, which enabled us to reassess and prioritize tasks through weekly meetings while maintaining an overall sprint goal. This approach facilitated a better adaptation to changing priorities and availability and helped maintain a continuous workflow despite external pressures. These regular

meetings and sprint planning sessions were crucial, providing a space to consider current progress, potential obstacles, and immediate next steps.

5.3.4 Choosing an algorithm

(Medium, Technical Related)
When choosing an algorithm, it was difficult to determine the best fit for the model. Initially, we had planned to choose the best algorithm by researching the best fit for our circumstances. This was useful and provided a starting point, but it did not provide the best option. In order to find the best algorithm, we had to compare several different algorithms using R-squared as the basis for evaluation. This proved very effective and resulted in us finding the best algorithm.

5.3.5 Refining the model

(Medium, Technical Related)
Refining the model posed several difficulties. These included selecting features and dealing with narrowing the dataset to relevant players by constraining age, league, and position. Initially deciding what features to include, we started with several simple features essential to determining the target variable. We tried to narrow age, assuming players on the boundaries of prime footballing years would act as outliers however, this preconceived notion was proven wrong through development. Different positions in football are valued differently, too, and because of this, the idea of creating a model for each position customized to the features most important to that position was the original plan for development; however, this changed upon results and further research. We can conclude that the impact position holds as a feature is more important than splitting the model into separate models per position. This provides a more holistic appraisal of a player because players are more versatile and malleable than simply capable of playing one position.

Chapter 6

Testing and Evaluation

6.1 Metrics

The extent of this project's development has led to a machine learning model that can predict a football player's transfer value based on scouting information. Several industry-standard metrics are used to evaluate a model.

6.1.1 Mean Absolute Variance (MAE)

MAE is the average of the absolute differences between the predictions and actual observations, providing a linear score representing the average error magnitude. It is particularly valuable in scenarios where each error contributes proportionately to the total amount of error. MAE gives a clear idea of what errors might cost without giving undue weight to large deviations, which can be critical in financial forecasting. MAE helps club managers understand the average error magnitude in monetary terms, offering a straightforward assessment of prediction accuracy that complements the more complex RMSE.
Formula:

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

[26]

6.1.2 Root mean square deviation (RMSE)

RMSE is the square root of the average of the squared differences between predicted values and observed values. It is a standard way to measure the error of a model in

predicting quantitative data. RMSE provides a scale-sensitive measure of prediction accuracy for predicting player transfer values. A lower RMSE indicates a model that more accurately predicts the transfer fees, which is crucial when these fees involve large sums of money. In practical terms, a lower RMSE enhances a team's ability to budget and plan for player acquisitions, minimizing financial risk.
Formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

[26]

6.1.3 Coefficient of Determination (R - Squared)

R-squared is a statistical measure representing the proportion of the variance in the dependent variable that is predictable from the independent variables. In the context of transfer value predictions, a higher R-squared value means the model explains a significant portion of the variance in transfer prices, indicative of a useful predictive model. A higher R-squared value assures stakeholders of the model's robustness and capability to perform reliably in various market conditions, thereby supporting strategic decisions based on the model's outputs.
Formula:

$$R - Squared = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

[27]

6.2 System Testing

The system testing of our machine learning model was meticulously conducted to assess its accuracy and reliability using real datasets. This phase involved a detailed and thorough setup that started with data preparation. The dataset used comprised scouting information and historical transfer values of football players for the year 2020 from the game Football Manager 2020. Attributes included age, position, and various performance metrics, which were crucial for the prediction model. To prepare this data for processing, we conducted several preprocessing steps that involved normalizing numerical data and encoding categorical variables to ensure compatibility and maximize prediction accuracy.

For model validation, we employed a K-Fold cross-validation approach with five splits. This method was carefully chosen to mitigate the risk of overfitting and to validate the model's performance across different subsets of data, providing a strong and reliable assessment of its predictive capabilities. The core of our testing involved the Gradient Boosting Regressor, configured with 100 estimators, a learning rate of 0.15, and a maximum depth of 3. These parameters were selected to balance between learning efficiency and predictive performance.

Each cross-validation fold generated predictions for the test set during the testing phase, after which we calculated several key performance metrics. These metrics included Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R-squared). MAE provided a straightforward measure of the average error magnitude, indicating the typical prediction error that could be expected by football clubs using the model. RMSE helped understand the variance in these errors, highlighting how significantly predictions could deviate from actual values. R-squared was used to measure the proportion of variance in the dependent variable that could be predicted from the independent variables, giving us an insight into how much of the movement in transfer values our model could explain.

In conclusion, this system testing comprehensively evaluated the predictive model using scientifically robust methods. It confirmed the model's efficacy in predicting football player transfer values and illuminated potential enhancements. This testing phase was crucial for demonstrating the model's capabilities and limitations, providing a balanced perspective on its applicability in the sports industry. The testing phase ensured clarity and maintained consistency throughout the analysis.

6.3 Results

This section provides a detailed analysis of the model's testing and evaluation results. The results are segmented into evaluations based on performance metrics and their implications for predictive accuracy and model reliability.

The model was tested using three key performance-related metrics mentioned in the previous section. For the first iteration of the model, we received these results while using features chosen based on intuition. These included age, height, finishing, technical ability, composure, decision-making, anticipation, on-the-ball ability, concentration, vision, and teamwork. These were the results.

- Iteration 1 Mean MAE - 6,593,061.02
- Iteration 1 Mean RMSE - 10,802,188,19

- Iteration 2 Mean R-squared - 32.99

This provided the starting point for the development of the model and we continued onto iteration 2. During this iteration the main focus was splitting the model based off of position.

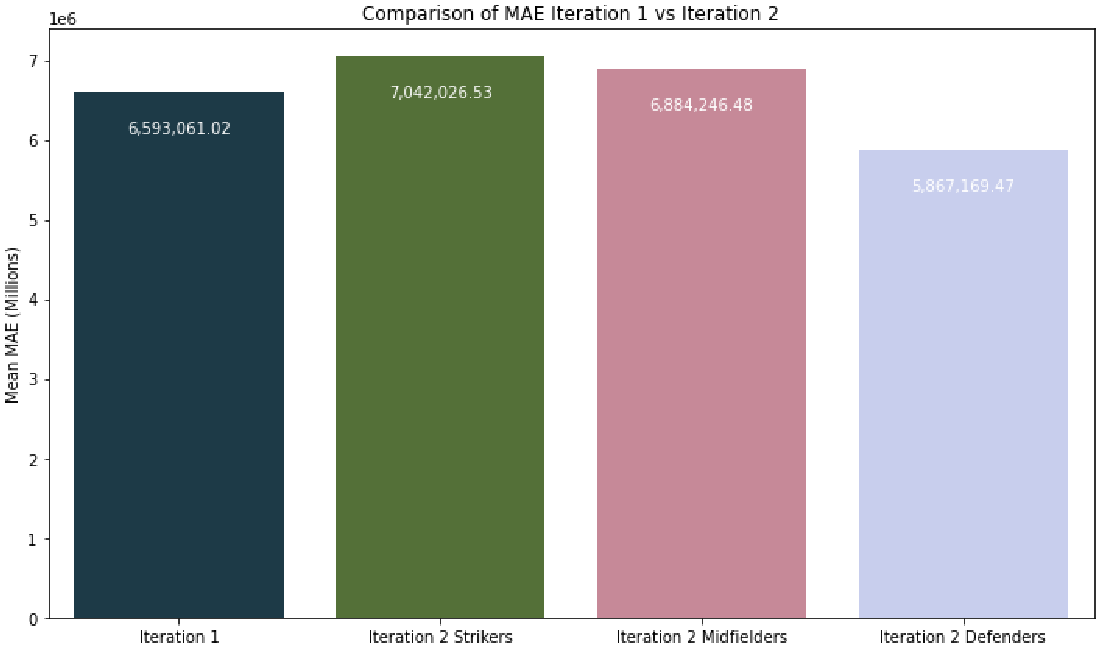


Figure 6.1: Iteration 1 vs Iteration 2 MAE

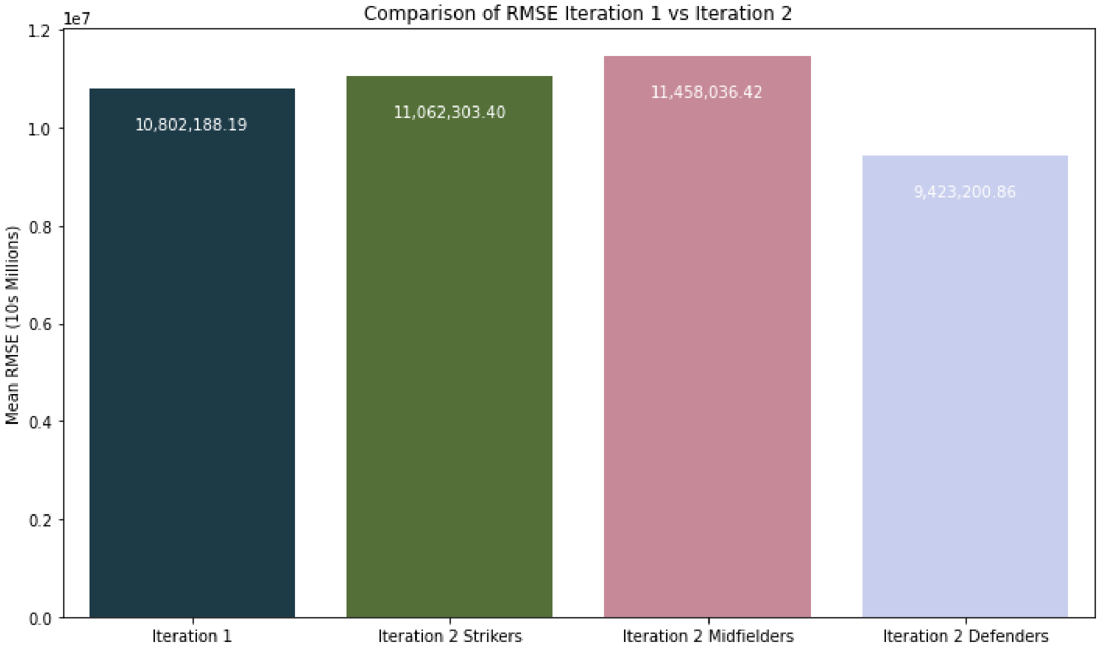


Figure 6.2: Iteration 1 vs Iteration 2 RMSE

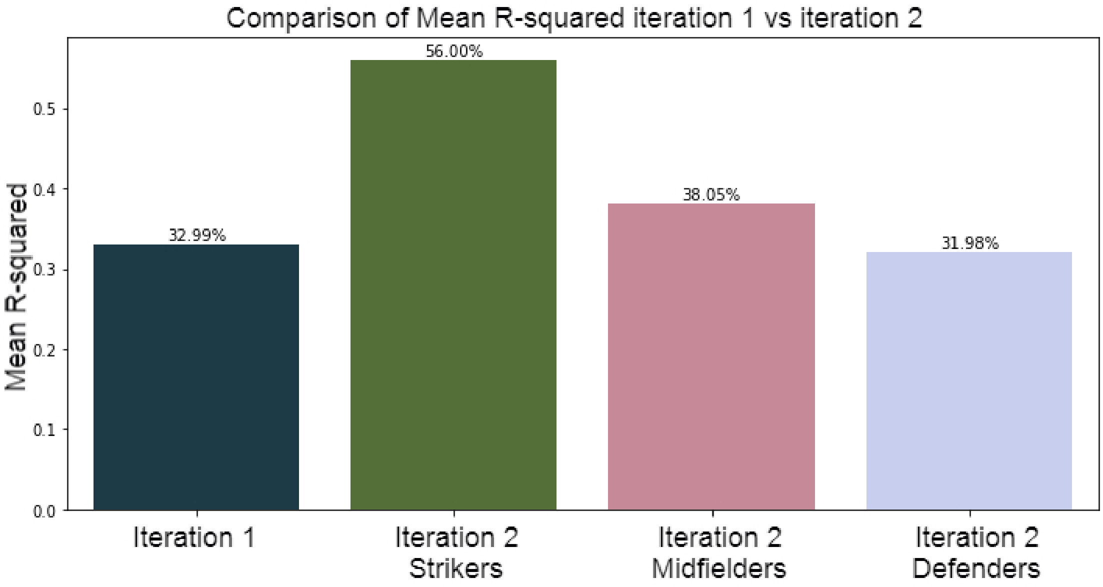


Figure 6.3: Iteration 1 vs Iteration 2 R-squared

We can see from this that the model was inaccurate, and the features selected were more closely aligned to Strikers than any other position. The RMSE and MAE of the models were quite similar, much more so than the R-squared, which suggests that the model consistently made errors of a similar magnitude across all predictions. The close values of RMSE and MAE indicate no extreme outliers significantly affecting the error metrics. However, the lower R-squared value indicates that despite this consistency, the model did not capture a substantial proportion of the variance in the dataset, suggesting a poor overall fit. This discrepancy highlights the importance of considering multiple metrics when evaluating model performance, as each provides unique insights into the accuracy and reliability of the predictive model.

In response to for the third iteration we evaluated different algorithms and decided to allow the model to use all features available and to remove and constraints applied to features.

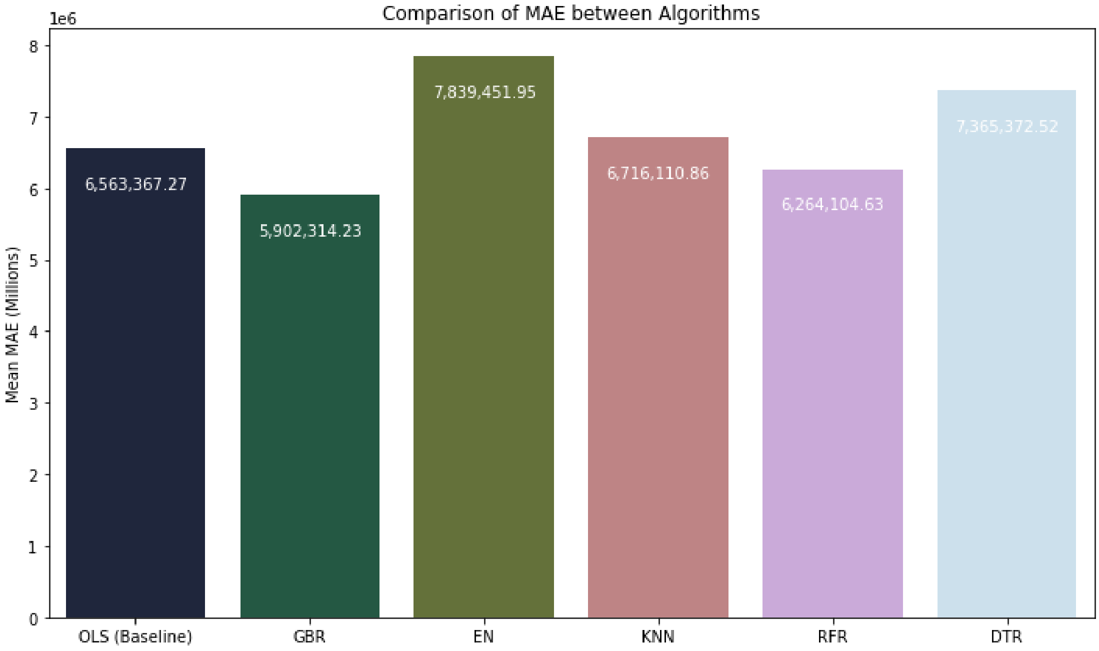


FIGURE 6.4: Comparison of MAE between Algorithms

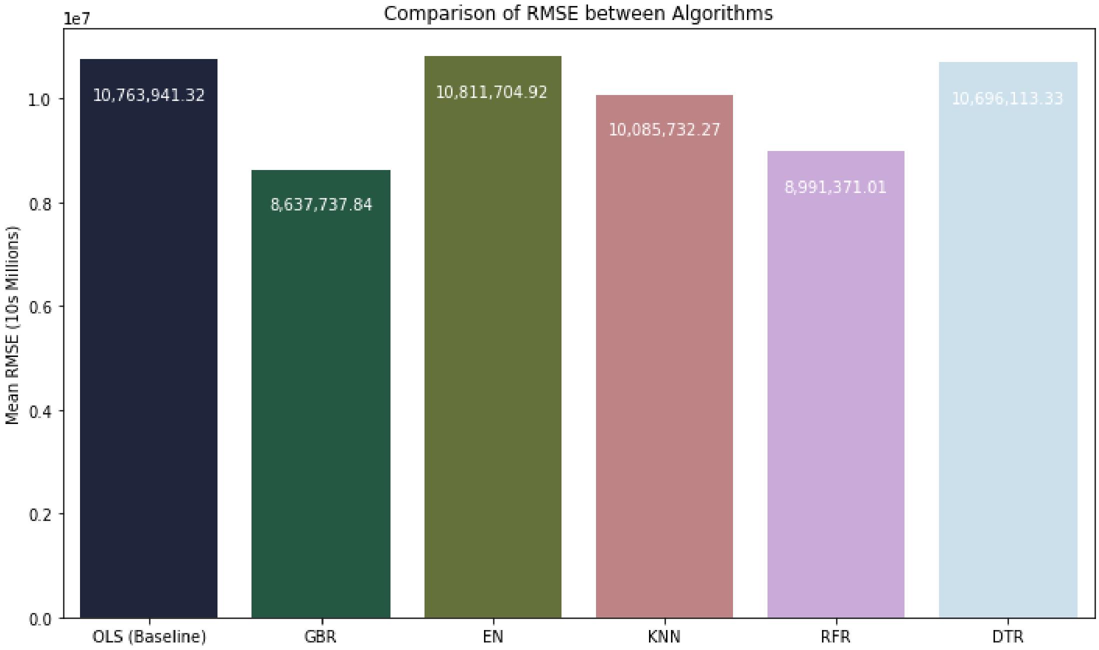


FIGURE 6.5: Comparison of RMSE between Algorithms

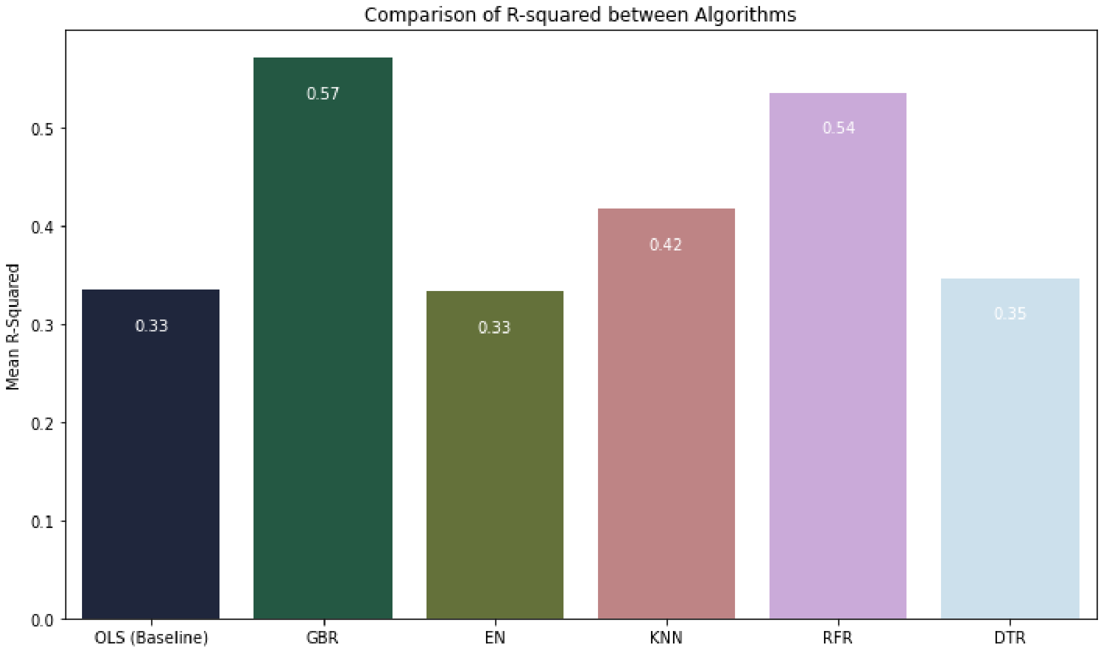


FIGURE 6.6: Comparison of R-squared between Algorithms

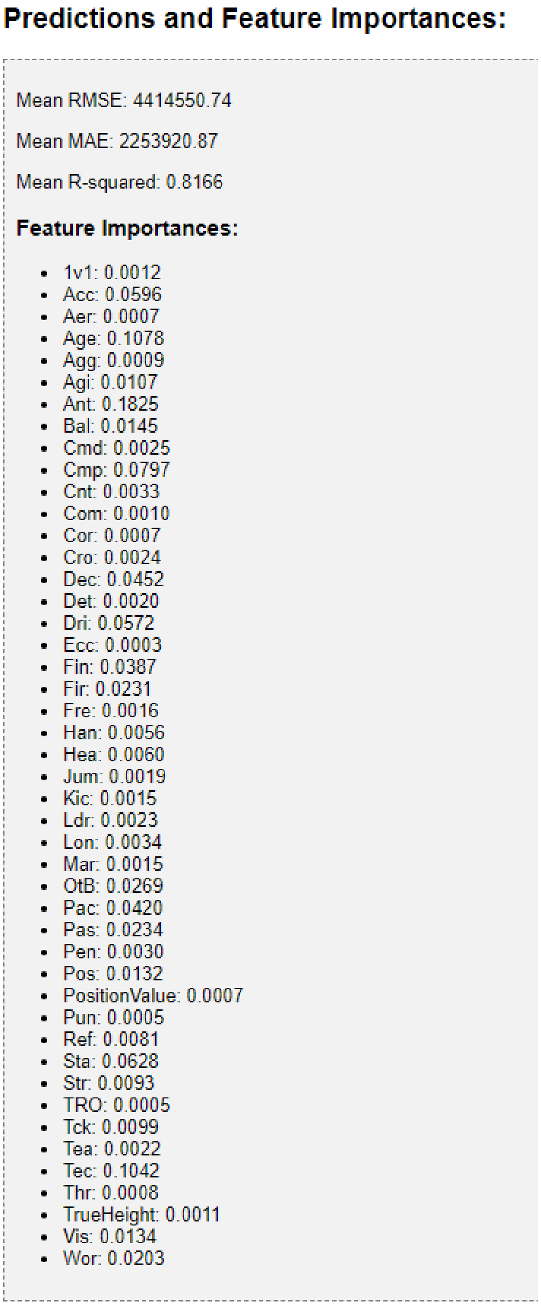
These results indicate that GBR is the optimal algorithm for this model. GBR has demonstrated superior performance across all metrics when compared to its counterparts. It achieved the lowest MAE at 5,902,314.23, which suggests that, on average, the absolute discrepancies between the predicted and actual values are minimal, enhancing the reliability of the model's predictions. The RMSE for GBR was recorded at 8,637,737.84. Being sensitive to large errors, this metric reaffirms that GBR effectively minimizes errors across the dataset. This is particularly important in the model where large errors can have significant consequences or indicate major prediction inaccuracies. Most impressively, GBR recorded the best R-squared value of 57%. This percentage indicates that approximately 57% of the variability in the dependent variable can be explained by the model powered by GBR. This robustness makes GBR a preferable choice for researchers and analysts seeking reliable predictive performance

For the final version of the model, the results for these metrics were the following:

- Iteration 3 Mean MAE - 2,253,920.87
- Iteration 3 Mean RMSE - 4,414,550.74
- Iteration 3 Mean R-squared - 81.66%

This model used 5420 players and 47 features including the target feature, transfer value.

In Figure 6.7, we can see these metrics and the importance of all the features. Several particularly impactful features include Age, Technical Ability (Tec) 0.1042, Anticipation



Predictions and Feature Importances:

Mean RMSE: 4414550.74

Mean MAE: 2253920.87

Mean R-squared: 0.8166

Feature Importances:

- 1v1: 0.0012
- Acc: 0.0596
- Aer: 0.0007
- Age: 0.1078
- Agg: 0.0009
- Agi: 0.0107
- Ant: 0.1825
- Bal: 0.0145
- Cmd: 0.0025
- Cmp: 0.0797
- Cnt: 0.0033
- Com: 0.0010
- Cor: 0.0007
- Cro: 0.0024
- Dec: 0.0452
- Det: 0.0020
- Dri: 0.0572
- Ecc: 0.0003
- Fin: 0.0387
- Fir: 0.0231
- Fre: 0.0016
- Han: 0.0056
- Hea: 0.0060
- Jum: 0.0019
- Kic: 0.0015
- Ldr: 0.0023
- Lon: 0.0034
- Mar: 0.0015
- OtB: 0.0269
- Pac: 0.0420
- Pas: 0.0234
- Pen: 0.0030
- Pos: 0.0132
- PositionValue: 0.0007
- Pun: 0.0005
- Ref: 0.0081
- Sta: 0.0628
- Str: 0.0093
- TRO: 0.0005
- Tck: 0.0099
- Tea: 0.0022
- Tec: 0.1042
- Thr: 0.0008
- TrueHeight: 0.0011
- Vis: 0.0134
- Wor: 0.0203

Figure 6.7: Importance of Feature

(Ant) 0.1825, Dribbling (Dri) 0.0572, Stamina (Sta) 0.0628, Composure (Cmp) 0.0797, and Determination (Det) 0.0502. Interestingly, Height (TrueHeight) 0.0011 does not have a significant impact. However, it is clear that the inclusion of these metrics, even if less significant, does benefit the model and is beneficial to include them.

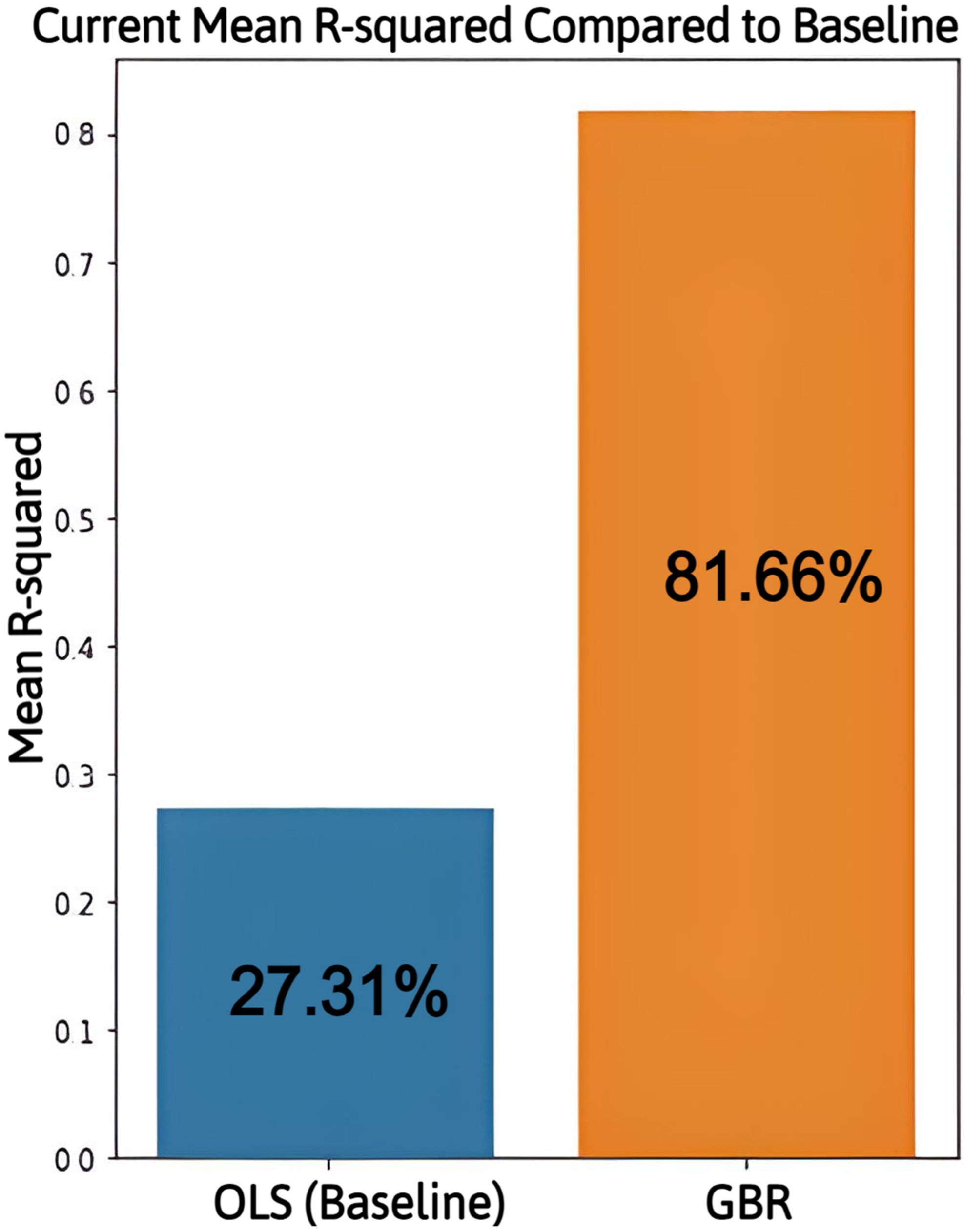


Figure 6.8: Comparative R-squared

Comparing the R-squared to the baseline model, OLS, with features and manipulation applied to the final iteration, we can see a stark difference in accuracy. OLS has an accuracy of 27.31%, while GBR has an accuracy of 81.66%. This difference has justified the use of GBR as the algorithm for the model.

For all models, we employed k-fold cross-validation with five folds to ensure the robustness and reliability of our predictive models. Specifically, we employed a cross-validation technique that divided the dataset into two distinct sets, which was 80% of the data was used for training, and the remaining 20% was reserved for testing. This 80:20 split is a widely recognized method in data science for balancing the need to train models on a substantial portion of the data while retaining a significant subset for testing the model's predictions. By teaching our models on the majority of the available data, we maximized the learning potential of the model, allowing it to capture relevant patterns and dynamics. Meanwhile, testing on the 20% holdout set provided a meaningful assessment of how well the model will perform on new, unseen data, thus validating its generalizability and effectiveness. This approach is essential for minimizing overfitting and ensuring that the model's performance metrics indicate its real-world utility.

Chapter 7

Discussion and Conclusions

7.1 Solution Review

The problem we aimed to provide a solution for was predicting the transfer value of a football player based on information about their abilities and body or, in other words, a scout report. This is the first significant problem that needs to be resolved to create a tool that can offer a nuanced and more holistic assessment of a player's level, which this project set as its goal. During the research phase of this project, we investigated the most essential elements that go into the valuation of a player. We discovered how players are valued differently for various reasons, such as attributes, position, technical prowess, and style of play.

Throughout the development phase of the project, we worked towards this goal. We successfully created a model that could quickly evaluate a football player based on some information. However, this is not a satisfactory solution to the issue. The first iteration of the model was not up to a level that was considered a viable solution. This begins the process of refining the model and improving our methods. Several iterations of the model were produced, and several different development avenues were explored. We arrived at our current iteration by investigating and exhausting all possible solutions. Our final model can explain over 80% of the variability of our target feature, which is transfer value. For the time allowed and invested into the development of this solution we are satisfied that it reaches the threshold of acceptable accuracy, marking a substantial step forward in applying machine learning techniques and evaluation to football.

This project highlighted the complexities involved in modeling human performance variables, which can be erratic. Addressing these complexities, our model incorporated a wide range of data inputs and underwent rigorous validation and testing. It is now at a stage with real-world use and could be applied in a professional setting. Moreover, the iterative nature of the model's development highlighted the importance of adaptability in predictive analytics. Each version of the model incorporated lessons learned from the previous one, whether it was adjusting the algorithms used, refining the dataset, or redefining the feature set. This adaptive approach was crucial in moving closer to an optimal solution that balances accuracy with practical usability.

7.2 Project Review

The project's original goal was to assess a football player's suitability for a transfer to a new club. Research quickly showed that this goal would not be easily obtainable. Rather than trying to develop a model that could predict suitability, which requires an incredible level of detailed information in many different areas outside of a footballer's abilities, the focus was shifted entirely onto the footballer and their footballing ability. This was done by using transfer value as the target variable. We learned from this that having a lofty goal is okay and helpful to strive towards, however, acknowledging and working on the steps required to achieve that goal is more important. What we would do differently is extensively research the suggested goal and uncover what large steps need to be taken to achieve it. In doing this, the work required to achieve the goal can be accessed, which provides the basis for setting an achievable goal for the project while striving towards the greater goal through the work.

Data is the cornerstone of any project that utilizes it. The entire project is shaped by the data it relies on. In our case, we were content with the data we had, even though it wasn't the intended dataset. We were fortunate that it was suitable for our needs. However, we recognize the importance of thoroughly reviewing the available data before embarking on any future projects that depend on a dataset. This includes assessing the quality, quantity, and any special requirements that the project may have. This thorough review process is crucial to ensure the data's suitability and to avoid any potential pitfalls in the project's execution.

When creating a model, the algorithm used for that model is essential. During the project, we discovered this by evaluating several algorithms for the best. We did this after already selecting an algorithm based on research. We learned from this that you must always test everything to observe the results and relationship. This is important so that you have complete knowledge of all variables within your work, be aware of what needs work, and have the ability to improve every area possible. After this experience, we know how important this process of investigating and evaluating your methodology is and would apply this to every area of the project to maintain a high quality of the work produced and exhaust all avenues for improvement.

Developing the model is a crucial phase in the project life cycle, as it serves as the convergence point for all project elements. This stage allowed us to fully grasp the significance of each component within the overall project framework. Effective planning and regular progress reviews played pivotal roles in streamlining this process. We maintained a steady pace throughout the development phase by utilizing weekly meetings to evaluate completed work and strategies for upcoming tasks. This consistent approach ensured that we were always prepared for the next steps with no delays caused by unresolved issues or uncertainties about future directions. The well-structured development plan we adopted was instrumental in avoiding bottlenecks and facilitating the optimal use of resources and time. As a result, this meticulous planning and execution strategy led to superior outcomes, underlining the effectiveness of our project management methodology.

The decision was made to develop a fronted for the project. An interactive web page where the model could be used and tweaked by the user. This was done to provide a platform for future work in the project. The belief is that by giving this simple interface with the project, if future work was done on this project, it would be much easier to continue on the work by providing another way to understand the work that has already been done. We were happy with this element of the project. However, there was consideration for this aspect in the project's research phase, it was never seen as necessary. This is incredibly important upon completion and reflection of development because it encapsulates the work done.

7.3 Conclusion

7.3.1 Primary Conclusion

The primary conclusion from this project is the successful creation and refinement of a predictive model capable of estimating a football player's transfer value based on their footballing abilities and physical attributes. This model could explain 80% of the variability in the transfer value. Given the complexities and difficulties involved in quantifying human performance in such a specific area, this is a significant achievement.

7.3.2 Secondary Conclusion

The project highlighted the importance of adaptability when conducting predictive analytics. The iterative development process structured through the use of agile was crucial for evaluating and planning the development of the project to progress toward our the goal. This approach allowed continuous improvement in the model's accuracy and usability and was not anchored by any assumption, conclusion, or method. This demonstrated that predictive modeling, especially in complex areas such as football or other sports, require a flexible and iterative methodology to achieve results.

7.4 Future Work

Discuss any proposals for completion of the project, or for enhancements, or for re-design of your solution or software. Enumerate all the things you would have wanted to do should you have more time to work on this project.

Future work on the project should focus on creating a tool that offers a more nuanced approach and provides a holistic assessment of a player's level. This could be done by continuing the development of the model created in this project or by creating a new model that provides insights into another area. This could be a model that deciphers the suitability of a player for a certain playstyle, a model that identifies the differences in the transfer market in different countries and can account for the different valuations in players across these markets or a model that predicts the future development player of a player.

If there was more time to continue developing this project, we would investigate more algorithms, in particular eXtreme Gradient Boosting, a variation of GBR, the algorithm used in the final version of the model. Additionally, we would like to expand the interactive webpage and turn this into a tool that could utilize these models. Without the addition of new models, we believe significant work can be done to improve it currently and that it is capable of growing alongside any model.

Bibliography

[1] M. H. Williams, *Nutrition for health, fitness and sport.*, 1999. [Online]. Available: https://www.cabdirect.org/cabdirect/abstract/19991410003

[2] G. B. G. Greco, A. Muscella and F. Fischetti, “Editorial: Physical stimulus- performance-adaptation: understanding the physiological relationship,” 2023. [Online]. Available: https://www.researchgate.net/publication/371367537_Editorial_Physical_stimulus-_performance-adaptation_understanding_the_physiological_relationship

[3] W. L. Kenney, J. H. Wilmore, and D. L. Costill, “Physiology of sport and exercise,” *Human Kinetics*, 2021. [Online]. Available: https://books.google.ie/books?hl=en&lr=&id=XoZGEAAAQBAJ&oi=fnd&pg=PP1&dq=importance+of+physiology+in+sports&ots=uQ2pmgBb3L&sig=kT9El6eGk4Tc65NdsoI3h5pTAto&redir_esc=y#v=onepage&q=importance%20of%20physiology%20in%20sports&f=false

[4] P. K. M. Mohr and J. Bangsbo, “Match performance of high-standard soccer players with special reference to development of fatigue,” *Journal of sports sciences*, 2003. [Online]. Available: https://www.researchgate.net/publication/10673250_Match_performance_of_high-standard_soccer_players_with_special_reference_to_development_of_fatigue/citation/download

[5] C. D. Bellefonds, “Why michael phelps has the perfect body for swimming,” *Biography*, 2020. [Online]. Available: https://www.biography.com/athletes/michael-phelp-perfect-body-swimming

[6] A. Kaiser, “Interesting facts about cristiano ronaldo,” *Footbalium*, 2023. [Online]. Available: https://footbalium.com/lifestyle/fact/382-interesting-facts-about-cristiano-ronaldo/

[7] R. D. H. Fullagar, S. Skorski and D. Hammes, “Sleep and athletic performance: The effects of sleep loss on exercise performance, and physiological and cognitive responses to exercise,” *Sports Medicine*, 2014. [Online]. Available: https://www.researchgate.net/publication/

266855811_Sleep_and_Athletic_Performance_The_Effects_of_Sleep_Loss_on_Exercise_Performance_and_Physiological_and_Cognitive_Responses_to_Exercise

[8] S. A. M. S. Brink, C. Visscher and J. Zwerver, "Monitoring stress and recovery: new insights for the prevention of injuries and illnesses in elite youth soccer players," *British Journal of Sports Medicine*, 2010. [Online]. Available: https://europepmc.org/article/med/20511621

[9] J. B. T. Koopmann, I. Faber and J. Schorer, "Assessing technical skills in talented youth athletes: A systematic review," *Sports Medicine*, 2020. [Online]. Available: https://link.springer.com/article/10.1007/s40279-020-01299-4

[10] C. J. A. d. S. M. G. Praça, V. V. Soares and I. T. da Costa, "Relationship between tactical and technical performance in youth soccer players," *Sensors*, 2015.

[11] M. R. S. Laborde and F. Dosseville., "Emotions and performance: Valuable insights from the sport domain," *Nova Publishers*, 2012.

[12] L. H. T. Woodman, P. Davis and N. Callow, "Emotions and sport performance: An exploration of happiness, hope, and anger," *Journal of Sport and Exercise Psychology*, 2009. [Online]. Available: https://www.researchgate.net/publication/24439424_Emotions_and_Sport_Performance_An_Exploration_of_Happiness_Hope_and_Anger

[13] R. Pollard, "Home advantage in football: A current review of an unsolved puzzle," *The Open Sports Sciences Journal*, 2008. [Online]. Available: https://www.researchgate.net/publication/228632270_Home_Advantage_in_Football_A_Current_Review_of_an_Unsolved_Puzzle

[14] L. Bransen and J. V. Haaren, "Player chemistry: Striving for a perfectly balanced soccer team," 2020. [Online]. Available: https://www.researchgate.net/publication/339675252_Player_Chemistry_Striving_for_a_Perfectly_Balanced_Soccer_Team

[15] P. W. R. G. D. Parnell, A. Bond and D. Cockayne, "Recruitment in elite football: a network approach," *European Sport Management Quarterly*, 2021. [Online]. Available: https://www.researchgate.net/publication/356772498_Recruitment_in_elite_football_a_network_approach

[16] M. L. D. Linke, D. Link, "Football-specific validity of tracab's optical video tracking systems," *PLOS One*, 2020. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230179#pone.0230179.ref015

[17] S. Sports, "Expected goals, expected assists, pressures, carries, high turnovers and more — advanced stats explained," *Sky*, 2023. [Online]. Available: https://www.skysports.com/football/news/11095/12829539/expected-goals-expected-assists-pressures-carries-high-turnovers-and-more-advanced-stats-explain

[18] T. Tureen and S. Olthof, ""estimated player impact" epi quantifying the effects of individual players on football (soccer) actions using hierarchical statistical models," *StatsBomb Conference*, 2022. [Online]. Available: https://www.researchgate.net/publication/363672422_Estimated_Player_Impact_EPI_Quantifying_the_effects_of_individual_players_on_football_soccer_actions_using_hierarchical_statistical_models

[19] P. A. Weinreich, "Data-based analysis in football transfer decisions," 2022.

[20] S. R. Department, "Big five - statistics facts," *Statista*, 2023.

[21] K. Stuart, "Why clubs are using football manager as a real-life scouting tool," *The Guardian*, 2014. [Online]. Available: https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool#:~:text=Indeed%2C%20in%202008%2C%20Everton%20signed,for%20life%20as%20a%20manager.

[22] XLSTAT, "Ordinary least squares regression," *XLSTAT*.

[23] K. Schmidheiny, "The multiple linear regression modelUnversity of Basel," *Unversity of Basel*, 2022.

[24] M. A. Al-Asadi and S. Tasdemır, "Predict the value of football players using fifa video game data and machine learning techniques," *IEEE Access*, vol. 10, pp. 22 631–22 645, 2022.

[25] J. Fernando, "R-squared: Definition, calculation formula, uses, and limitations," *Investopedia*, 2023. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp

[26] P. M, "A comprehensive introduction to evaluating regression models," *Analytics Vidhya*, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

[27] N. University, "Coefficient of determination, r-squared." [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html