When Moneyball Meets the Beautiful Game: A Predictive Analytics Approach to Exploring Key Drivers for Soccer Player Valuation

Yisheng Li

Master of Science in Management

Submitted in partial fulfillment of the requirements for the degree of
Master of Science in Management (Operations and Information Systems)

Goodman School of Business, Brock University
St. Catharines, Ontario

Dedicated to my maternal grandmother

Abstract

To measure the market value of a professional soccer (i.e., association football) player is of great interest to soccer clubs. Several gaps emerge from the existing soccer transfer market research. Economics literature only tests the underlying hypotheses between a player's market value or wage and a few economic factors. Finance literature provides very theoretical pricing frameworks. Sports science literature uncovers numerous pertinent attributes and skills but gives limited insights into valuation practice. The overarching research question of this work is: *what are the key drivers of player valuation in the soccer transfer market?* To lay the theoretical foundations of player valuation, this work synthesizes the literature in market efficiency and equilibrium conditions, pricing theories and risk premium, and sports science. Predictive analytics is the primary methodology in conjunction with open-source data and exploratory analysis. Several machine learning algorithms are evaluated based on the trade-offs between predictive accuracy and model interpretability. XGBoost, the best model for player valuation, yields the lowest RMSE and the highest adjusted R^2. SHAP values identify the most important features in the best model both at a collective level and at an individual level. This work shows a handful of fundamental economic and risk factors have more substantial effect on player valuation than a large number of sports science factors. Within sports science factors, general physiological and psychological attributes appear to be more important than soccer-specific skills. Theoretically, this work proposes a conceptual framework for soccer player valuation that unifies sports business research and sports science research. Empirically, the predictive analytics methodology deepens our understanding of the value drivers of soccer players. Practically, this work enhances transparency and interpretability in the valuation process and could be extended into a player recommender framework for talent scouting. In summary, this work has demonstrated that the application of analytics can improve decision-making efficiency in player acquisition and profitability of soccer clubs.

Keywords: *Moneyball, sports analytics, player valuation, predictive modeling, interpretable machine learning*

Acknowledgement

Over the course of my time in the MSc program, many people have given their blessing to me. I want to express my deepest gratitude to all of them. Thank you!

First and foremost, I am immensely grateful for my intrepid supervisor, Dr. Anteneh Ayanso. My thesis would otherwise be deemed not “mainstream” and its completion would not have been possible without his vision. His mentorship plays a pivotal role in managing the scope and writing style of this thesis. He motivates me to stay proactive and stay ahead of the schedule. Besides, his advice reinforces my cherished conviction that true merits will eventually triumph.

I want to extend my heartfelt thanks to my extraordinary supervisory committee members, Dr. Shuai Yuan and Dr. Martin Kusy. Dr. Shuai’s meticulous and thoughtful feedback is indispensable to the quality of this thesis. Dr. Kusy’s big picture thinking prompts me to relentlessly think hard and think outside the box. I truly appreciate all the comments the external examiner Dr. Mehmet Begen gives. Thank Dr. Tianyu Guan for chairing my thesis defense.

A very special homage to my sports analytics course instructor Brian Burke at George Mason University. Brian was instrumental in sparking my epistemic curiosity.

Thank many operations and information systems faculty members at Goodman who generously offered me TA opportunities, which has been essential for a foreign student who lives paycheck to paycheck and does not have much support from some conventional sources.

Last but not least, I would like to pay tribute to my parents, for all their love and sponsorship. During my mother’s college years, there was an apparent ceiling for women to continue education. My mother left big shoes to fill. It is my distinct privilege to carry on her legacy. I dedicate this thesis to my maternal grandmother, a career educator who always treats people with uncompromising kindness.

Table of Contents

Table of Tables

Table of Figures

Chapter 1 Introduction

1.1 Background

Professional sports bear a great deal of resemblance to corporate business. A modern sport club is no longer just a team but rather a sophisticated business organization wherein games are a vehicle of entertainment for which other businesses produce an imperfect substitute (Rottenberg, 2000). It has loosely coupling units, weakly cohesive and fragmented tasks, ambiguity of preference and competing objectives (Flegl et al., 2018). A plethora of companies have long hinged on data to enhance decision-making and streamline nearly every aspect of their business. Sports clubs, by and large, are lagging behind in this trend (Davenport, 2014a). Hunch and gut feelings became a deeply ingrained tenet of decision-making in sports that leaves some puzzles unresolved. What is the "true value" of a player? How can it be measured in a disinterested fashion? A few trailblazers invented an unorthodox approach, later knows as Moneyball, to fill the void (Davenport, 2007). Grounded in this approach, this thesis draws on insights from relevant business theories and sports science to address critical questions in soccer player valuation and employs data exploratory analysis, predictive analytics, and model explanation methods to augment soccer club's decision-making in player acquisition.

Moneyball is a phrase coined by the renowned non-fiction writer Michael Lewis in his bestseller *Moneyball: The Art of Winning an Unfair Game*. It is most commonly referred to the analytical, evidence-based strategy adopted by the low-budget Oakland Athletics to recruit undervalued baseball players at the dawn of the 21st century. In the context of baseball, Moneyball is synonymous with Sabermetrics. As originally defined by the legendary Bill James in 1980, often dubbed the "founding father" of the intelligent use of baseball statistics, Sabermetrics, at a very high-level, is the quest for objective baseball knowledge primarily via statistical analysis ("A Guide to Sabermetric Research," n.d.). Baseball is well suited to statistical modeling. Conversely, modeling many other sports is a daunting task from a computational standpoint (Davenport, 2014b). The buzzword analytics has risen to prominence. Academics and practitioners use sports analytics as an umbrella term concerning the extensive use of data and quantitative methods to gain a competitive edge in and beyond the sports arena (Davenport, 2014b). A formal and restrictive definition of sports analytics is "the management of structured historical data, the application of predictive analytic models that utilize that data, and the use of information systems to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play" (Alamar & Mehrotra, 2011a, para. 2).

A sports analytics taxonomy characterizes three major categories (Cokins et al., 2016): individual sports (e.g., golf, tennis), team sports (e.g., soccer, basketball), and league sports management (e.g., National Football League, National Basketball Association) that coordinates teams or individuals. The three categories subsume front-office "business-side", back-office "team-side" and other topics related to sports and its societal impact. Some minor categories of the two sides are: 1) sports business operations analytics, including scheduling, fan promotions, digital marketing strategy and dynamic ticket pricing; 2) player and team performance analytics such as recruiting and scouting players; 3) health, nutrition and injury prevention analytics. The prestigious annual *MIT Sloan Sports Analytics Conference* co-founded by Daryl Morey, the president of basketball operations of the Philadelphia 76ers, offers a venue not only for practitioners and executives but also for researchers and aficionados to discuss the landscape of analytics in the global sports industry.

Sports clubs have incentives to pursue a virtuous circle of sporting success and financial prosperity. Efficient clubs are capable of minimizing the acquisition costs of players and maximizing the athletic performance of their squad such as improving their standing in a league or winning champions. A bottleneck is the vacuum of theoretical rationales coupled with analytics techniques that are crucial for ramping up the capacity to identify cost-effective playing talents (Davenport, 2014b; Gerrard, 2017). Accurate player profiling and recruiting, therefore, serve the best interest of a club. Over the past two decades, the economic dynamics of the labor market in professional sports has profoundly changed, and the capital pouring into the market has skyrocketed. In soccer, the landmark Bosman ruling in 1995 favored greater employment mobility for players, prompted an influx of international players in the domestic leagues thereafter and ushered the labor market of soccer (*aka* transfer market), into an increasingly globalized talent pool. The transfer market *per se* is a multi-million lucrative industry that features blockbuster deals and multi-million payrolls. The gross global spending on transfer fees has more than tripled over the past 15 years and exceeded €10 billion in 2019 due to the meteoric growth of club revenues (Poli et al., 2020). A handful of cash rich clubs can afford exorbitant transfer fees to land a marquee player, deliberately or inadvertently inflating the price tags on players. The most expensive soccer transfer record hitherto was set by the move of Neymar from FC Barcelona to Paris Saint-Germain for €222 million in August 2017 (Conn, 2017). Kuper et al. (2015) argued that the transfer market is often inefficient given the weak association between large spending and commensurate rewards. Although the new arrival of a *galácticos* (superstar in Spanish) instantly galvanizes fans, that star may have already peaked or only done well in the international tournaments.

Gerrard (2017) spotted a “capability gap” that the capability to collect sports data far exceeds the capability to make sense of that data. Soccer clubs desperately crave analytics prowess, as corporations are vying for the best analytics minds. The Moneyball philosophy is entrenched in Liverpool’s organizational culture. A high-profile instance is Liverpool’s analytics team which consists of a small cadre of academics and statisticians. The team created a proprietary database to track the progress of more than 100,000 players worldwide, recommending which bargains Liverpool should pick up, and then how the new signings might be used (Schoenfeld, 2019). Those backroom analytics specialists play an unsung role in Liverpool's restoration as a continental and domestic powerhouse. In lieu of splashing money on expensive established players, Liverpool has relentlessly hunted underrated players and turned them into stars (e.g., Mo Salah, Sadio Mane, and Roberto Firmino). Grooming young talents and then selling them for a profit has been a business model for some clubs in top-tier leagues to stay solvent. Moneyball is by no means limited to baseball or soccer. Wayne Winston, an emeritus professor of operations and decision technologies at Indiana University, had been advising the Dallas Mavericks vis-à-vis lineup selection and free agent market strategy. He was instrumental in Mavericks’ NBA champion journey in 2011. Based on analytics, Wayne recommended that the rising prospect Devin Harris be traded for a much older veteran Jason Kidd. This trade sparked a backlash from fans and media pundits, since it seemed to be at odds with the conventional wisdom that Kidd had passed his prime when Mavericks signed him. A metric called “clutch impact factor” underpinned this trade decision: when Kidd was on the court, the score seemed to move in his team’s favor. Proven by historical data, he had a track record of consistently using his basketball finesse to elevate teammates (LinkedIn Learning, 2013).

1.2 Motivation

As González (2008) elegantly put, modern soccer grew out of migration from rural to urban areas and the alienation of the new proletariat from the old bourgeois, as a mass phenomenon that was eerily intertwined with Marxism and a legacy of the accelerated industrialization in the 19th century. Soccer is a universal language people use as an expression of *esprit de corps*. Desmond Morris, an observant British zoologist, drew parallels in his book entitled *The Soccer Tribe*: soccer can act as a social drug, in which hordes of fans (tribal warriors) march through the streets to the stadium (the Great Temple) chanting praise for their team and mockery songs for the opponent (enemy). The Brazilian legendary footballer Pelé used the Portuguese phrase 'O Jogo Bonito' (meaning *'The Beautiful Game'*) to name his autobiography *My life and the beautiful game*.[1] Moneyball captures the tipping point in the soccer data revolution when clubs embrace disruptive innovations. The transformation of the global soccer industry through analytical decision-making is in progress. This thesis is motivated to explore the degree to which Moneyball represents transferable knowledge in soccer, since analytics in the transfer market is an intriguing, relevant and relatively uncharted research theme in terms of managerial implications, modeling hurdles and emerging techniques.

First, sports analytics has nontrivial managerial implications for business value creation. Like any other profit-oriented firms, a quintessential question for soccer clubs is to determine the value of their roster of players (intangible assets), as players are clubs' most important human capital and their acquisition costs and wages account for an enormous proportion of clubs' total spending. The most consequential decisions the club leadership takes revolve around player acquisition and retention. Personnel decisions that are solely reliant on intuition and common sense could squander a massive amount of money (Kuper & Szymanski, 2012). The use of data and analytics, as effectively demonstrated in other industries, can help the club leadership make sound and long-term value investments (Sierksma, 2006). The benefits of applying analytics in the transfer market include more precise approximation of the valuation of intangible assets, insights into the expenses and costs of the player's employment and the margins of the potential transaction (Majewski & Majewska, 2017). Davenport (2014a) argued that the adoption of analytics in sports can teach general managers, head coaches, and players to align leadership at multiple levels, focus on the human dimension, and work within a broader ecosystem.

Second, soccer is the holy grail of sport analytics since it is often perceived as resistant to the pull of solid number-crunching, which constitutes a thorny challenge for researchers. Striking and fielding team sports such as baseball and basketball are most amenable to data analytics techniques due to their discrete nature. Every game can be dissected into units. In baseball a natural unit is a pitch or an at bat. In basketball it is a 24-second shot clock. Individual playing contributions are discernible in such sports. The high degree of separability is conducive to the systematic exploitation of player performance data. At the other end of the spectrum stands fluid, stochastic invasion sports like soccer in which replicating the Moneyball formula is considerably more problematic. Invasion team sports are a type of sports such that a group of players cooperate to move an object (e.g., a ball) to a designated target defended by opponents (e.g., between goalposts). In soccer, there is no apparent unit other than the 45-minute halftime mark. Changes in possession could be units, but a team's single possession could last several minutes long. Gerrard (2007) detailed several dimensions of complexity that reside in invasion team

[1] Soccer, football, and association football are synonyms herein and appear interchangeably.

sports: player actions are often concerted actions such as several players joining together to block a shot from an opponent in soccer; soccer also has a continuous transition between offensive and defensive plays, unlike American football which is segmented. Barring technical hurdles, organizational obstacles regarding data governance and technologies must be overcome (Rein & Memmert, 2016). Alamar and Mehrotra (2012) noted a cultural conflict between proponents of Moneyball and traditionalists who have first-hand experience in the sport. the most enduring barrier to the Moneyball approach being generalized is the cultural barrier. A shift to model-driven analysis would amount to a seismic paradigm change. Sports data scientists typically come from a technical background such as computer science or statistics but do not necessarily possess deep appreciation of the game. They need to articulate their findings in a plain, jargon-free language, as the vast majority of players, coaches, managers and executives are analytics laymen. The club leadership needs an analytics vision and changes the organization to unleash analytics' potential.

Third, the data deluge makes soccer an ideal laboratory in which advanced data analytics techniques are experimented. Soccer analytics marries domain expertise with computational toolkits of data science in the era of big data. Historically, human experts were enlisted by clubs to observe and assess potential target players through paper-and-pencil methods, often in person. Such human-based scouting is unable to scale and often subject to biases. The "availability heuristic" suggests that the more easily people recall certain information, the more credit they give it (Tversky & Kahneman, 1973, p. 207). For example, scouts are inclined to privilege short-term performances over long-term trends and overweigh their own experience. Sports clubs fall short of decision support tools to recalibrate intuitive value judgement. Nowadays, video libraries and performance data start to proliferate through IoT (Internet of Things) techniques. Some crown jewels of sports data problems have become tractable by virtue of advanced analytics in conjunction with big data. The adolescence of soccer data paves the way for a more dispassionate investigation of what makes a sensible transfer. As a nascent discipline, sports business research through the lens of analytics is still in its infancy, with many groundbreaking discoveries transpiring in real time. This is especially true in soccer. Many cutting-edge analytical endeavors are published online (blogs, social media, etc.) and are not peer-reviewed. Aspiring researchers are embarking on a quest of tacit knowledge pertaining to decision-making. With all in-person scouting being banned due to the COVID-19 pandemic, and given all the uncertainty in the future, clubs which accommodate new technologies ahead of rivals will be able to convert an analytical edge into a competitive edge at scale.

1.3 Research Questions and Contributions

Motivated by these opportunities and implications, this thesis focuses on soccer player valuation. Valuation generally can be seen as an assessment of the estimate worth of a commodity. The connotation of valuation varies subtly from discipline to discipline (e.g., accounting, economics or finance). To keep the research scope manageable, this study narrows player valuation down to the process to estimate realistic market values of players. In other words, how much are they worth in the transfer market? A related field is player evaluation which rates players by their performance or attributes but seldom estimates their monetary values. Clubs usually appeal to expert opinions on player valuation. However, what criteria or metrics experts use remains a conundrum. Some criteria or metrics turn out to be cognitive blinders because they are more of an art than a science. Therefore, the main objective of this research is to unravel the pillars of players valuation, and, at a granular level, the features underpinning these pillars, using analytics as a means. The overarching research question is:

What are the key drivers of player valuation in the soccer transfer market?

To address this question, this work draws on the literature in market efficiency and equilibrium conditions, pricing theories and risk premium, and sport science, and attempt to answer the following specific questions:

1) *How do economic and risk factors affect player valuation in the transfer market?*
2) *How do physiological attributes, psychological attributes and soccer skills affect player valuation in the transfer market?*

This work then proposes an integrative conceptual framework for player valuation. To empirically validate the proposed framework, this work leverages publicly available transfer market data and statistical and machine-learning-based techniques and algorithms. There is scant academic work applying predictive analytics to player valuation, since current predictive studies in soccer concentrate on score forecasting or players' and team performances (Tunaru et al., 2005). This work takes a predictive analytics approach to distinguishing the signal from the noise in the transfer market. The signal corresponds to the variations of the key drivers of a valuation, whilst the noise is variation omitted by researchers either due to complexity or the protracted unavailability of information (Franks et al., 2016).

Theoretically, this thesis advances the frontier of business analytics research in sport management by deepening our understanding of the key drivers of player valuation and empowering clubs with more efficient decision support tools for talent scouting. The main contributions of this thesis include:

- a conceptual framework for soccer player valuation through a unified lens of established theories from economics, finance and sport science;
- an empirical validation of the proposed framework by harnessing a state-of-the-art predictive modeling technique coupled with open-source soccer data;
- an appropriate trade-off between predictive accuracy and model interpretability through the exploratory data analysis and a novel interpretable machine learning method.

This thesis also contributes practical and social benefits through an accurate and interpretable prediction model of player valuation that has a long-term effect of building transparency and trust in the notoriously opaque transfer market. With those models at disposal, club executives and managers, who are grappling with the optimal allocation of the budget for player acquisition, would gain operational and economic benefits. In addition, guided by the data-driven insights of this work, clubs will have the opportunity to acquire a player who has unfulfilled potential at a reasonable transfer fee (e.g., an undervalued player), or be well informed of how much they would overpay for a player. As an outcome, clubs are to channel limited resources to prospective players who have certain traits clubs are longing for. This will be particularly relevant for clubs in their efforts to overcome the financial predicaments that may linger in the post-pandemic era. Although clubs are the main beneficiary of the findings of this research, players can also use those findings to set and manage a proper expectation of themselves. To sum up, this thesis has important implications for club profitability and efficient decision-making in player acquisition.

The rest of this thesis is structured as follows. Chapter 2 provides a comprehensive and systematic literature review that gives background information for the transfer market and the regulatory environment and synthesizes the relevant theories and findings. Chapter 3 proposes a conceptual framework of soccer player valuation. Chapter 4 details the data sources and methodologies. Chapter 5 interprets the key findings of empirical data analysis, including the most important features in the best predictive model. Chapter 6 discusses the results of data analysis and some implications. Finally, Chapter 7 encapsulates the conclusion and limitations of this work, pointing out a few directions for future research.

Chapter 2 Literature Review

This chapter provides a systematic review on the literature pertinent to player valuation. Section 2.1 emphasizes some key concepts of labor market in the context of professional sports and explains the measures of transfer procedures. Section 2.2 examines the theoretical foundations of player valuation by surveying a stream of related work, encompassing economics, finance and sports science.

2.1 Context - Labor Market in Professional Sports

In North American sports leagues (e.g., NFL, NBA), professional athletes are normally “traded” for other athletes. That is, a team should offer its own players or draft picks in exchange for players from another team. The NFL has a draft system whereby every team’s selection of a college talent is determined by a reverse order of teams final standing in the previous season. The process of determining which new players the various teams choose is called a draft (Florke & Ecker, 2003). Likewise, the NBA has a centralized draft lottery system to hire rookies. This lottery system has introduced a degree of randomness such that the team with the worst record is only guaranteed to receive a high probability of being bestowed the first pick. It refrains teams from intentionally losing more games to increase their odds of getting a higher pick, since NBA superstars are overwhelmingly high draft picks. In principle, draft systems serve as checks and balances that ward off rich teams draining talent reservoirs and breeding a winner-takes-all market. Dynasties wax and wane. Thus, no franchise can uphold hegemony for decades.

By contrast, a transfer in professional soccer occurs when a player moves to a new club, specifically referring to the transfer of a player's registration to a new club (Swanepoel & Swanepoel, 2016). Majewski and Majewska (2017) shed lights on what is transferred precisely - “footballer’s performance rights”. That is, a professional football player on every division is obligated to be registered in a national football association. Gerrard (2014) added that the football player’s performance right is the exclusive right to field the player in games. From a financial perspective, performance rights are an intangible asset of the club. Cash settlements are incurred in the form of transfer fees or other payments, providing a club has acquired a player from another club during his contract. Advocates of the transfer market contend that clubs are entitled to recoup their investment in training and development of a player (Coluccia et al., 2018). A permanent transfer is implicitly referred to as “transfer”, while a temporary transfer is referred to as “loan”. The estimation of loan fees is fundamentally different from that of transfer fees. My research focuses on permanent transfer, excluding free transfer and player swap.

Transfer fee and market value are conceptually different yet comparable concepts (Herm et al., 2014), although both frequently appear in academic papers. A transfer fee is the actual amount of money a club has to pay for a player’s performance rights. It is the final value placed upon the player by his incumbent club which the purchasing club agrees to pay. This is an unusual element of the transfer market, since researchers have no such information in many other settings where workers are not allowed to be “bought” and “sold”. A proxy for that transfer fee is colloquially called the market value. A player's market value is “an estimate of the amount of money a club would be willing to pay in order to make [an] athlete sign a contract, independent of an actual transaction” (Herm et al., 2014, p. 484). Transfer fees can be higher and lower than market values due to the length of the remaining contract, strategic reasons (e.g., undermining a rival club by buying its key player at a price higher than usual), or the bargaining power of the buying and selling club.

Business researchers often reckon a general question: what is the relationship between the value of its workers and the firm valuation? Ployhart et al. (2014, p. 373) developed a definition of human capital resources: “Individual or unit-level capacities based on individual knowledge, skills, abilities, and other characteristics (KSAOs) that are accessible for unit-relevant purposes”. As argued by Ployhart and his colleagues “Human capital resources based on interactive or causal complementarities have greater opportunities for enhancing performance and generating competitive advantage than resources in isolation” (Ployhart et al., 2014, p. 385). For football clubs listed in a stock exchange, their market value does not primarily derive from the value of its facilities or other capital stocks. Instead, it is created by implicit ‘human capital stock price’ of players – their football knowledge and skills. Their transfer fees and salaries warrant that their services should be recognized as a human resource accounting asset or intellectual capital. The club’s stock price is the result of an aggregate function of the players ‘stock prices’. KPMG (2020) ranks football clubs by enterprise value (EV) that is calculated as the sum of the market value of the owner’s equity plus total debt, less cash and cash equivalents, regardless of the capital structure used to finance its operations.

The transfer market is too global to implement draft mechanisms, but it does have an intricate regulatory regime to promote competition. Since 2010, stringent financial fair play regulations have been in place both by UEFA (Union of European Football Associations), the governing body for association football in Europe, and by individual leagues for the purpose of “improving the overall financial health of European football” (UEFA, 2019). Pursuant to the rules, clubs are discouraged to spend astronomical amounts of money and must balance their books in three years. Compliance failures will result in sanctions such as fines, transfer embargos and even temporary expulsion from European competitions. Salary cap is another remedy for inequality. Every club has a certain maximum quota on players’ salaries. Exceeding that limit will be penalized by hefty luxury taxes. To circumvent the penalties for breaching the financial fair play rules, clubs do not account for the cost of players in the form of net spend. Instead, they apply an accounting method called, player amortization, that evenly splits the transfer fee and wage paid throughout the economic lifespan (i.e., the contract duration) of the player (Amir & Livne, 2005). As a result, the annual total acquisition cost reported on the balance sheets of the club equals the amortized transfer fee plus annual wages.

Multimillion transfer deals undergo the multifold procedures. Negotiation over the price of players has been institutionalized and followed due diligence. Before any party approaches the negotiation table, the club will have spent months or even years scouting the target as well as viable alternatives. Though the coach is part of the recruitment process, it is usually the “director of football” or senior executive who has the final say. The sporting director model is a precautious measure to shield clubs from ever-increasing financial risks associated with their sporting decisions and has been fruitful at many clubs across various soccer leagues. If the target is currently under contract at another club, the direct liaison between him and the buyer club is strictly prohibited. Third-party shareholders (investors, funds, economic rights) could further complicate the liaison. The buyer club must formally make a bid for the player. The two clubs work out a mutually agreeable transfer fee that the vendor will receive. Once the vendor grants the permission, the transfer will proceed with the negotiation phase: the buyer will offer a contract to the player mostly via his agent. Agents broker a deal in the hope of driving up their clients’ market value and sometimes are rewarded with decent commission fees. Resourceful agents boast a robust network of clubs and even wield enormous power of orchestrating and facilitating a transfer.

The important terms of the contract players and agents iron out include, *inter alia*, wage, contract duration, conditional terms such as performance bonus, compensation for using player’s image right and release clause. The vendor might contingent future payments on the player winning any major trophy or making another lucrative move. If the player is content with the contract terms and passes medical examinations, the deal can be officially inked. Club counsels and immigration law lawyers deal with legal affairs. For membership-based clubs like Barcelona and Real Madrid, all registered members are *de facto* shareholders. However, they are not stakeholders: they are not consulted on strategic or operational decisions which are delegated to the board of directors and the manager or coach. Although scouting, bidding and negotiation could be initiated anytime throughout an entire calendar year, players are only permitted to join a new club from their current club during prescribed time windows. The FIFA Regulations on the Status and Transfer of Players (FIFA, 2020a, p. 13) states that players may only be registered during one of the two annual transfer windows as per the leagues. The first window shall normally open after the completion of the season, last no more than twelve weeks and end prior to the new season. The second one normally commences in the middle of the season and may not exceed four weeks. In a race against time, clubs work around the clock to push a transaction through before the deadline.

2.2 Theoretical Foundations

This thesis has two major building blocks as the theoretical foundations of player valuation. The first block encompasses market efficiency and equilibrium conditions from the economics literature (Section 2.2.1) as well as pricing theories and risk premium from the finance literature (Section 2.2.2). This provides the theoretical lens to explore fundamental economic and risk factors as key drivers of player valuation. The second block is grounded in sports science that enriches our understanding of physiological and psychological attributes as well as soccer skills that give a critical piece of information for player valuation (Section 2.2.3).

2.2.1 Market Efficiency and Equilibrium Conditions

Market efficiency theory refers to the degree to which market prices incorporate all available, relevant information (Fama, 1970). Market efficiency theory stresses the paramount importance of exploring as many valid constructs as possible to achieve accurate player valuation. Implicit discussions of market efficiency are common in the transfer market, as signings are judged according to whether the contracts represent players’ “fundamental values”. In an informationally efficient transfer market, market values incorporate and reflect all relevant information. Therefore, players should be sold for their exact valuation not more or less. This is where theory contradicts the reality of the football world. Information like past performance is not necessarily a reliable predictor of future success (Allen, 2018). Clubs at the top of the football hierarchies are often based their valuations on anchors (Sæbø & Hvattum, 2019). Anchoring is a behavioral economics theory that when making a valuation, humans have a cognitive bias where they rely too heavily on an initial piece of information to guide subsequent judgements. Real transfer negotiations have been going on behind the scenes and some key details remain undisclosed. One of the ramifications is the presence of undervalued or overvalued players. The following sections only discuss publicly available information.

Market equilibrium is the state at which supply and demand curves intersect, and as a result, market prices are stabilized. The transfer market is a labor-intensive market, where human resources strongly affect organizational performance (Wright, Smart & McMahan, 1995). A seminal paper claimed that the ultimate objective of football clubs is to maximize utility (Sloane, 1971). The term utility describes the measurement and satisfaction that a consumer obtains from any good or service (Taussig, 2013, p. 124). In the soccer context, acquisition and exchange of players by clubs aim at boosting team performances and the chance of winning, thereby maximizing utility (Carmichael & Thomas, 1993). Clubs bid for the players' services, and in equilibrium, the final bid price of a player can be thought of as a function of the valuation of winning attributes of a player (Rastogi & Deodhar, 2009). Footballers are semi-homogenous, yet they have vastly different characteristics which make their market values vary immensely. Football is a highly specialized profession. Nurturing a young prodigy is extraordinarily hard. It is also not unusual to witness many burgeoning footballers failing to blossom. Under the umbrella of market equilibrium, the transfer market is governed by the basic law of supply and demand. The demand for talents often spikes while there is invariably a shortage of supply, especially at the high end, where only clubs with large cash pile place a serious bid. This unique fabric is manifested by the price premium paid by clubs —the excess price paid over the baseline price that is justified by the expected economic value of a player. Interestingly, Rao and Bergen (1992) illuminated that a price premium paid by quality-conscious buyers is in fact an economically rational endeavor to secure promised level of quality for experience products. The same holds true for the transfer market. Clubs sometimes knowingly pay a price that is higher than what is justified by the relative quality of the player. Such footballers usually have scarce attributes and cannot be cultivated on a large scale. In a nutshell, market equilibrium accounts for why and how economic determinants like footedness, position, historical data, nationality and superstar status drive supply and demand curves.

The majority of professional soccer players are right-footed (Yorke, 2019), which makes left-footed and two-footed (ambipedal) players a scarce resource. Fry, Galanos and Posso (2014) found a premium of being a left-footed player. Bryson et al. (2013) pointed the finger at evidence of a substantial salary premium for two-footed ability, *ceteris paribus*. The observable variation in transfer fees can be explained by the similar variables that also affect remuneration (Frick, 2007). Two-footedness, as a rare trait, increases market values in two ways. First, having two feet of roughly equal strength means a broad shooting angle and quick reaction under complex situations. The adeptness to use both feet makes the player a nightmare to defend against and afford him transient opportunities that one-footed players are unable to seize. Therefore, some well-rounded forwards exhibit a more balanced distribution of goals between both feet. Secondly, two-footed players tend to be more versatile since theoretically they could fill more positions. For instance, a two-footed winger can excel in either flank, whereas a right-footed winger could only feel comfortable with one side and perform poorly in another side. The club has the luxury to use a two-footed player in several positions on the pitch and this positional utility may generate a return to transfer fees. The substantial premium for two-footed players is consistent with the proposition that two-footedness diversifies players' tactical value (Bryson et al, 2013). In short, two-footedness is tangentially advantageous gift that are rewarded in the transfer market.

Transfer fees appear to vary by position and the degree of specialization. Attacking players are highly sought after and hence can command a higher price. Unsurprisingly, elite forwards and attacking midfielders who specialize in creating goals or assists dominate the list of the most expensive players,[2] as their contributions are most conspicuous on the pitch. Defensive midfielders and goalkeepers are systematically undervalued. Some researchers identified a

[2] https://www.transfermarkt.com/statistik/transferrekorde

renumeration premium earned by midfielders and forwards relative to defenders (Frick, 2007; Bryson et al., 2013). Goalkeepers are the most specialized player who are not apt for any other position. Yam (2019) outlined two caveats of goalkeeper evaluation: their dependence on team's defensive strength as well as on opponent's offensive strength; and the scarcity of goalkeeper actions. Midfielders are the least specialized (hence the most versatile) players who can play several positions and assume different tactical roles. Left-footed footballers who can play left back and left center back are in demand. Left-footed players are purported to perform more naturally in both positions. It is difficult for clubs to find qualified left-footed players to occupy such positions given the scant supply of left-footed players. Pappalardo et al. (2019) quantified the notion of versatility as a player's flexibility to switch position or role from match to match. The added flexibility would be helpful while crafting lineups and the transfer market may organically gravitate to this value. Thus, a position-aware valuation framework incorporating the specificity of each position is long overdue.

Certain nationalities may lead to a price premium. This is due to the regulatory environment where international transfer certificates and work permit applications affect a transfer. Clubs have limited options for players who meet the requirements in regard to nationality. For example, UK-born players of the English Premier League (EPL) have a notable premium partly due to the protection of British labor law and a specific stipulation that eight of any club's 25-man first-team squad must have spent at least three years at an English or Welsh academy before their 21st birthday. The Italian first division, Serie A, imposes a maximum limit of three 'foreign' (non-EU) players per match-day squad. UEFA mandates that international players who have spent at least three seasons between 15 and 21 years old in the employer club could count as club-trained (i.e., "homegrown") players. Those players are at a premium (Berg, 2011). Nationality also has a cultural dimension. Perciballi (2011) argued that expatriate footballers from different ethnicities will experience impulsive degrees of cultural assimilation. Kuper and Szymanski (2012) in their *Soccernomics* cited anecdotal evidence that English clubs had a long-standing preference for Scandinavian players given their attainment of English proficiency and cool climate adaptability. The most frequent migrant route of footballers is from Brazil to Portugal, which reflects the shared language and colonial history. Pedace (2008) found that South American players tended to be overpaid. This pattern presumably associates with perceptions and precedents that South American players, particularly Brazilians, are "naturally" more talented.

Historical data have been consistently studied by soccer economics research. For example, appearances in domestic leagues, in the European leagues and on the national team all have a positive effect on transfer fees (Frick, 2007). Understandably, the number of times a player is substituted during a season has a negative effect on market values (Lehmann, 2000). "Minutes played" is discriminative variable that can translate to other metrics (Franks et al., 2016). Unlike basketball, soccer is a low-scoring game in which goals and assists are rare events and sometime happen in a haphazard way. Crude descriptive statistics like goal and assist do not fully mirror a player's true value. Soccer economics studies could benefit from advanced performance metrics (Sloane, 2015). This necessitates the need to conjure up omnibus metrics like expected goals and expected assists. Both are inspired by the expected value theory. Expected value of a random variable is a generalization of the weighted average over a large number of experiments. Expected goals (xG) is the probability that a shot will end up with a goal.[3] Likewise, expected assists (xA) measures the likelihood that a given pass will be

[3] For example, an xG of 0 means no chance to score whatsoever, while an xG of 1 is an actual goal.

converted into a goal assist. In other words, xA gives an indication of how many assists a player should have had. Assists are a common measurement of creativity. In addition, xA assigns a fair level of credit to the player who makes the pass regardless of the final result of that pass (goal or no goal). Club scouts have used xG and xA to evaluate a target player's attacking efficiency (Rathke, 2017).

The "superstar phenomenon" is defined by a landmark paper to be one "wherein relatively small numbers of people earn enormous amounts of money and dominate the activities in which they engage" (Rosen, 1981, p. 845). In soccer, only a handful of players are widely recognized as superstars and can impact on a franchise that transcends the club itself. Patnaik et al. (2019) observed that this phenomenon wields disproportionately high positive influence on those "superstar" players' transfer fee, albeit they might be just marginally better than corresponding players. Adler (1985) in his influential paper indicated that the positive network externalities of popularity set superstars apart from equally talented performers. Garcia-del-Barrio and Pujol (2007) elaborated that popularity of footballers does not entirely stems from in-field contributions. In a similar vein, Herm et al. (2014) found that player popularity can be differentiated from players' intrinsic skills. Although the magnitude of the talent-related popularity is plausible, Franck and Nüesch (2012) noted that the nonperformance-related celebrity status of a player measured by press publicity increase the market values of soccer stars. Sports should be viewed from a broad media and entertainment perspective (Kobielus, 2014)). Media equate charismatic players with glamorous Hollywood stars. Clubs are intentionally paying more for popular players due to their global commercial desirability and crowd-pulling power. In return, players' popularity presumably enhances the economic profitability of their clubs through tickets, merchandise sales, sponsorship, commercial deals, image rights, and broadcast revenues. Besides, being under contract with an elite club as such is emblematic of the class of the player. Such players might reap benefit from brand recognition in the form of remuneration or transfer fees. Shapiro (1983) showed that reputation facilitates a price premium; hence, reputation building can be considered as an investment good.
Players who have character issues and become embroiled in a flurry of scandals would seriously damage their market value. A club with a poor reputation (e.g., lower average league position) is less appealing to players.

Equilibrium can be extended from the basic law of supply and demand to bargaining. Bargaining theory and game theoretical framework (Nash games) anatomize ubiquitous bargaining in the transfer market (Carmichael & Thomas, 1993). Contract is the result of hard bargaining between players and clubs. A strand of literature concerned contract length being a major determinant of the transfer fees (Patnaik et al., 2019; Carmichael, Forrest, & Simmons, 1999). The 'Monti-regime' stipulates a maximum contract duration of 5 years. Frick (2011) pointed out that players sometimes become less motivated due to huge financial stimulus offered by guaranteed multi-year massive contracts. The number of remaining year(s) in a contract is a particularly delicate issue. Clubs face a dilemma when one of their key players has only one year remaining in his contract and declines to renew it. The Bosman ruling forbids the incumbent club from commanding a transfer fee when the player's contract has expired. Clubs could either sell this player as soon as possible at a transfer fee that is pronouncedly lower than his market values; or let him walk away from the club for free at the expiration of his contract—a calamitous economic loss for the club.

Nash equilibrium sheds light on such a competitive environment where players have greater bargaining leverage to hold their owners to ransom (Anonymous, 2017). A Nash equilibrium comprises a set of strategies, one for each agent, such that no agent can improve marginal gains by altering its course of action given what it predicts the other agents would do (Massey & Thaler, 2013). It is natural to contemplate the talent acquisitions in the transfer market as a noncooperative game, since clubs decide independently how many players to recruit and how much to pay them, subject to the rules and bylaws of the leagues. Imbalance of bargaining power exists among clubs. Not all buyers are price takers—some clubs exert more influence than others (Szymanski, 2004). Swanepoel and Swanepoel (2016) discovered a strong correlation between the bargaining power of the buying club and transfer fees. Carmichael and Thomas (1993) examined the transfer fee within the Nash bargaining theory—the greater the player is, the stronger the bargaining position of the selling club is. Bargaining power can be operationalized by the club ranking (e.g., domestic champion, promotion or relegation) and transfer fees and wages spending (Frick, 2007). Mourao (2016) also found that the efficiency of transfer inflows can be significantly influenced by a long sports history and the presence in the season's UEFA Champions League or UEFA Europa League.

Auction theory is an applied branch of game theory to describe the bidding process in the transfer market. The sellers always have a bidding value on their prospective player, and buyers have to match the value. Sellers sometimes place a price on their players insofar as it is much higher than their market value through release clause, in an attempt to deter potential buyers from launching a hostile bid for the player the seller intends to retain. However, if a buyer activates the release clause, the seller will have no choice but to approve the transfer and no auction will happen. Otherwise, the market runs on a first price bidding auction, where the bidding price is not hidden from other buyers. Different clubs can submit multiple bids and the seller will only accept the bid if the price matches or exceed the proposed value of the player. Thus, the amount of the transfer fee would be the outcome of "a bargaining process" (Rottenberg, 2000).

The aforementioned studies have laid groundwork by highlighting relevant economic theories. The market equilibrium theory tests the underlying (casual) hypotheses between market value or salary and a few economic factors (e.g., age, nationality, playing time). The market efficiency theory suggests an inclusion of relevant information from wide-ranging theoretical lenses. However, these studies only utilize a small number of economic factors from basic demographic and coarse performance data. Those factors, though essential, do not capture all the relevant information in the transfer market and are insufficient in determining the accurate market value of a player. Economic theories do not explicitly provide a proper pricing framework of players. Theoretical lenses beyond economics are needed to broaden our understandings of player valuation.

2.2.2 Pricing Theories and Risk Premium

Finance is a school of thought that gives pricing theories and incorporates risks of player valuation. Hedonic pricing theory reinforces the central point that more information contributes to more efficient valuation and fills the gap of actual pricing left by market equilibrium. Rosen (1974, p. 34) first presented it in his paper—"Hedonic prices are defined as the implicit prices of attributes and revealed to economic agents from observed prices of differentiated products and the specific number of characteristics associated with them." There are two underlying hypotheses of this pricing model: 1) the observed market price of a good or service reflects the sum of implicit (aka hedonic) prices for its utility generating attributes (Rosen, 1974); a good or service can be treated as a repertoire of attributes that differentiate it from other goods or services

(Rastogi & Deodhar, 2009). 2) hedonic pricing model further postulates that the pricing factors of a complex product can be decomposed by internal factors of the good being sold and external factors affecting it. The services of a footballer are a differentiated product for which the prices (transfer fees) are disclosed, and characteristics (e.g., economic determinants) can be evaluated. Hence, its price is nothing but the summation of the hedonic prices of all embodied attributes. Gerrard (2001) used this pricing model to identify player characteristics indicative of their future market value.

Before investing in a player, the club needs to weigh in on the hazards and devise risk-hedging strategies, as it would do for other investment projects, such as building a stadium. Club's athletic success, financial might, historical status and attendance are positively related to the club's degree of risk aversion (Carmichael & Thomas, 1993). Sometimes the transfer fee paid is not subsequently vindicated by the net gains accruing to the purchasing club, which resembles stock market bubbles. For clubs with substantial amounts of money at stake, age and injury are two most silent risks. Age is a source of *ex-ante* risks because its relationship with market values is predictable. Professional soccer players have a relatively short career span. A consensus in the research community is that the age of 27 is a watershed in players' career development when their athletic performance starts progressively to decrease, irrespective of the sport practiced, and they finally retire in their mid-thirties (Stambulova, Stephan & Jäphag, 2007). Perennial players like Roger Federer and Michael Jordan are among a few exceptions. Majewski and Majewska (2017) drew an analogy between a player's career trajectory and the life cycle of a financial derivative, using the option price theory. They discretized a player's market value development throughout his career into 4 phases: 1) the introduction phase of a linear trend; 2) the growth phase of an exponential trend; 3) the stabilization phase of a logarithmic trend; and 4) the decline phase of a power trend. In the first two phases, age reflects potential and players appreciate with value added through training and match experience. Their summarization suggests that soccer players, over the course of their career, are analogous to investment assets that incrementally appreciate and return a future dividend, reach a plateau and after entering the final phase, depreciate due to the ageing process and injury. It has implications for modeling market value: age, typically as a continuous variable, may also be discretized as a categorical variable that has a succession of development stages.

While risk factors like age are the circumstances under which coaches and managers largely anticipate the outcomes, *ex-post* risks like injuries sometimes emerge in an unforeseen and involuntary way (Degli et al., 2015). It is appropriate to valuate a player by using a portfolio of options on him in an uncertain environment with uncertain cashflow and specific risks (Coluccia et al., 2018). Tunaru et al. proposed an option pricing framework for player valuation in which uncertainties like injury events were explicitly included (Tunaru et al., 2005). The theory of decision under uncertainties help forecast the probability of injury occurrence in the future and develop mitigation strategies. If a player rarely gets injured or has a resilient physique (i.e., making a swift and thorough recovery), he will be more likely to be assigned a higher valuation in the transfer market than a comparable player who is nevertheless prone to injury. When comparing players on a par with each other, clubs should favor players that have adequate stamina and robustness, focus on players that can bounce back from the rigors of traveling, training and playing, and look at the rate of games played as well as the volume. Discipline and adaptation and are two slightly less severe risks. The number of yellow and red cards a player received are a measure of disciplinary issues, as an excessively high number could pose a liability. A red card and the ensuing suspension would significantly reduce win probabilities.

Taylor and Giannantonio (1993, p. 474) defined organizational adaptation as "the process through which an individual comes to understand the values, abilities, expected behaviors and social knowledge that are essential for assuming an organizational role and for participating as a member". This definition also applies to adaptation in a soccer club (Al-Madi et al., 2016). New player's struggle to align in style of play culture or training drills could endanger club's investment on him. His form might decline, or the team (even with him) might not fare well.

Table 2-1 is a summary of the factors that have been examined by economics or finance literature to explain player valuation, including their theoretical foundation and target variables. It is important to note that sports science knowledge is needed to fully explain why some of these factors affect player valuation. Age has effect on a player's core attributes such as endurance and agility. To measure the injury risk is a study area of sports science. Although business researchers have made strides in selecting and extracting variables related to player valuation (Berg, 2011; Patnaik et al., 2019; Rottenberg, n.d.; Carmichael & Thomas, 1993; Majewski & Majewska, 2017; Tunaru et al., 2005), finance research on player valuation, like its close cousin economics, does not include sufficient sports science factors. New insights into player valuation may emerge in the intersection of business research and more sports specific aspects.

Table 2-1: Economic and Risk Factors Examined in Related Studies

Domain	**Theoretical Foundation**	**Factor**	**Target Variable**	**Reference**
Economics	Market Equilibrium	Footedness	Salary	Fry et al., 2014; Bryson et al., 2013
		Position	Salary	Frick, 2007; Bryson et al., 2013
		Nationality	Market Value	Berg, 2011
		Popularity	Market Value	Patnaik et al., 2019
	Nash Bargaining Theory	Bargaining Power	Market Value & Salary	Carmichael & Thomas, 1993; Rottenberg, n.d.
		Contract Length	Market Value	Patnaik et al., 2019; Carmichael, Forrest, & Simmons, 1999
Finance	Option Price Theory	Age	Market Value	Majewski & Majewska, 2017
		Injury	Market Value	Tunaru et al., 2005

2.2.3 Sports Science – Physiological and Psychological Attributes and Soccer Skills

Personnel selection in complex organizations involves defining not only social factors but also physical and psychological characteristics and measuring individual attributes (Flegl et al., 2018). Sports science is a collection of knowledge, theories, and research methods in sports psychology, sports health, and sports informatics (Röthig et al. 2003; Baca 2014). Many physiological and psychological attributes and soccer skills have been extensively studied by sports science research (Reilly et al., 2000; Martin, 2016; Williams & Reilly, n.d.; Martin & Miller, 2016; Williams, 2000; Ali, 2011). Reilly et al. (2000) investigated talent identification in soccer players by a multidisciplinary test battery that embraced physiological, psychological and soccer-specific performance measures, including somatotype (body shape), body size, anaerobic power (speed), aerobic power (endurance), technical skill, anticipation, task, ego orientation and the like. Malcolm Gladwell (2011) in his revelatory book *Outliers: The Story of Success* observed that even the birth month could impact physiological attributes. A disproportionate number of soccer players were born towards the first three months of the year. Hirose (2009) documented that the distribution of birth month among Japanese adolescent soccer players was skewed such that numbers were greatest in Quarter 1 and smallest in Quarter 4. Gladwell explained that age differences of less than 12 months in children could make a substantial difference in individual biological maturation. Young players enter youth academies and junior leagues based on a January 1st age cutoff, so those who have the earlier birthdays have a head start that snowballs into more coaching attention and play time to mature. Youngsters born earlier in the selection year are privileged compared to their cohort born later in the same year, a phenomenon called the relative age effect (Barnsley et al., 1992). A conclusive remark is that "individual performance thresholds are determined by our genetic make-up, and training can be defined as the process by which genetic potential is realized" (Tucker & Collins, 2012, p. 555). Sport psychology is the scientific study of an athlete's thoughts and behavior as they pertain to sport (Martin, 2016). Psychological measures mainly test personality such as self-confidence, anxiety-control, motivation and concentration (Williams & Reilly, 2000). Sport confidence, a major psychological construct, is defined as the belief an individual possesses about his or her ability to be successful in a sport (Martin, 2015). Anxiety can be both facilitative and debilitative (Burton & Naylor, 1997). Seasoned players cope with anxiety better than less experienced players. Deterioration in performance is often attributed to a lack of motivation, just as winning against a superior opponent is attributed to strong motivation (Vallerand, 2004).

Physiological and psychological attributes are general measurements, largely independent of what type of sports athletes practice. An individual may possess remarkable physiological and psychological capacity but is unable to compete at the highest levels of a particular sport due to the lack of sport-specific skills (Martin & Miller, 2016). This is not to downplay any of those attributes. Rather, the combination of the abovementioned characteristics and soccer-specific characteristics perform well in terms of talent identification (Huijgen et al., 2014). Performance attributes can be categorized as motor skills, cognitive, and perceptual skills (Ali, 2011). The motor skills to control, pass, dribble and shoot the ball at goal are building blocks of the soccer player. The ability to score goals is the most valued skill. Dribbling the ball past opposing players is one of the most eye-catching and entertaining soccer events. The ability to dribble, therefore, is widely accepted as a distinguishing feature of gifted players. Accurately passing the ball to a teammate and the act of heading a ball are also fundamental aspects. Ali (2011) gave an insight into the relationship between skills and techniques: the skill is a learnt ability to select and perform the appropriate technique and therefore the cognitive component, in the form of decision making, is a fundamental skill. Simply put, good decision-making means players know the right timing to shoot, pass or dribble. Perceptual and cognitive skills are less visible to spectators and receive less credit than motor skills, including attention, anticipation, decision-making, game intelligence and creative thinking (Williams & Reilly, 2000). The Nobel economics laureate Daniel Kahneman (2012) invented the concept of "System 1"—that is, a system of the human mind operates quickly and heuristically, generating complex patterns without much effort or a sense of voluntary control. For instance, when a player controls the ball and decides to make a pass, the basic elements of the situation the player will need to perceive are which teammate to pass to and opponents who may interfere. Many psychological processes (primarily situational awareness) may be activated in making a successful pass besides perceptual ones, including attention, memory, decision making and action (Jones, 2005). the capability to search and exploit space on the pitch makes elite players distinct from ordinary players. To illustrate this point, Fernández et al. (2019, p. 1) cited the words of the late Dutch legend Johan Cruyff: “it is statistically proven that players actually have the ball 3 minutes on average. So, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you are a good player or not.” Intelligent off-ball movement patterns require capacity to process spatiotemporal information in real-time before taking the optimal course of action. “*Raumdeuter*” (roughly translating into space investigator or interpreter) is the German phrase to describe a distinctive playing style of the footballers who mainly rely astute appreciation of space to chip in with goals and assists (e.g., Thomas Müller). Sports psychologist Zoe Wimshurst concluded that the superb awareness of constantly scanning the space beyond the defenders and anticipating in advance where the ball is going greatly contribute to player performance (McDowall, 2011). Dr. Barbara Sattler inferred that left-footed players appear to have a better perception of space than their right-footed counterparts (DW Kick off!, 2019). Left-footed players tend to unleash more creativity since they make unpredictable or counterintuitive moves for defenders to handle. Although not every left-hander is also a left-footer, footedness and handedness are related in most people (Tran & Voracek, 2016). Denny and O’Sullivan (2007) suggested that left-handed people may be cleverer than otherwise similar right-handed people citing a correlation between left handedness and IQ. Left-footedness is a natural gift, but two-footedness could be learned through deliberate practice. In general, for both left-footed and two-footed players, their physical dexterity may be associated with greater mental dexterity.

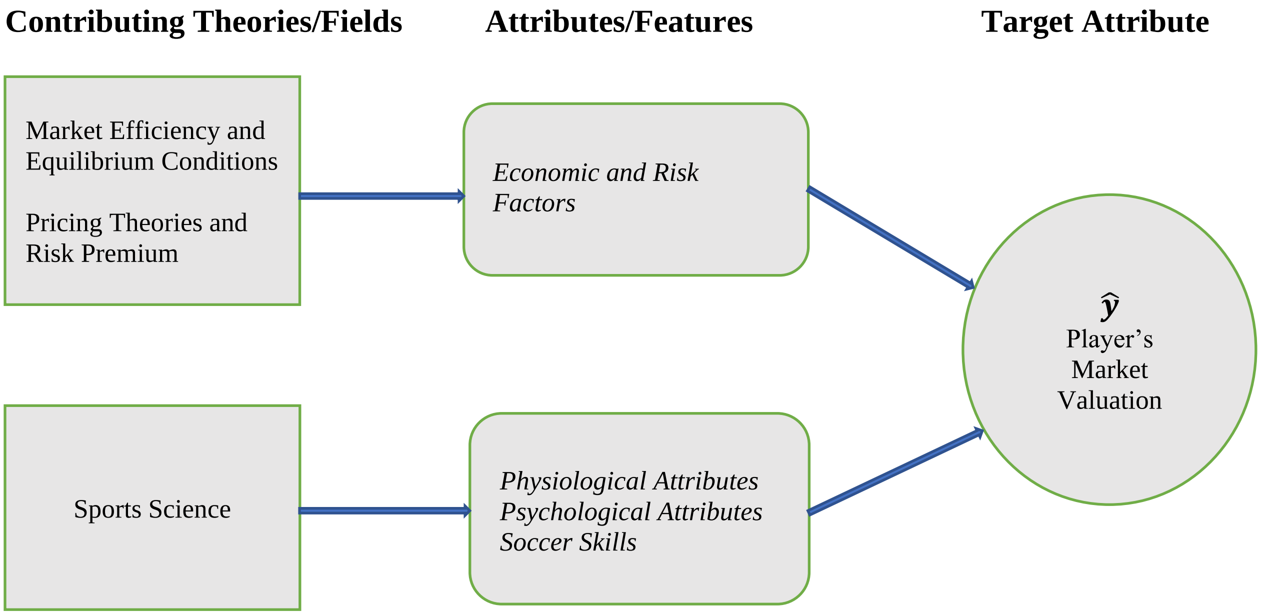
The soccer terminology “form” refers to either a player or a team’s recent performance. A central premise is that form influences future success. Form can be modeled as latent variables estimated from various ratings. Although rating and ranking are used interchangeably, they are distinct terms. A numerical score designated or assigned to a specified player or team refers to its rating. The term ranking refers to the order (rank) by which the list of players or teams are organized. Typically, sports rankings are determined by wins versus losses and an indicator of the market value of the athlete (Martin, 2016). Opta Index, a performance rating, has been proven to be reliable and valid predictors for player valuation (Tunaru et al., 2015). Kharrat et al. (2019) used ratings to investigate the efficacy of recruitment decisions. Many stakeholders in the sports industry keep abreast of rating and ranking since these figures provide a convenient way to judge player quality, predict performance, and assign market value. It is worth to note the imparity of collective ability across teams and leagues as acknowledged by UEFA. Some leagues (e.g., EPL) are more competitive on average than others (e.g., Dutch Eredivisie). Two players at practically same level would perform differently if one is in a stronger league whilst another is in a weaker league. A bias could be induced such that players in weak leagues receive inflated ratings.

Chapter 2 has probed the conceptual background for the selection of attributes and synthesize scientific observations that complement intuitive value judgements. Nevertheless, the sports science factors compiled by Chapter 2 are not exhaustive, nor are the economic or risk factors. While a body of sports science research pays close attention to many related attributes and skills, it neglects to put those attributes and skills into a broad business perspective. The interconnection between many sports science factors and player valuation has not been well established. Business research elucidates how player valuation fits into the framework of standard economics and pricing theories but has not fully exploited sport science research.

Chapter 3 develops a conceptual framework of soccer player valuation that unifies theoretical foundations of value drivers from this chapter and applies to a wide range of variables.

Chapter 3 Conceptual Framework of Soccer Player Valuation

Soccer player valuation is a multifaceted process involving knowledge acquisition and representation from different aspects of the transfer market environment. This includes not only player-specific insights, but also knowledge related to external factors and the specific market conditions. Traditionally, the focus of sport science has been on player evaluation, with limited insights into player valuation. In attempting to fill this gap, this research proposes a conceptual framework for soccer player valuation (Figure 3-1). To the best of my knowledge, the proposed conceptual framework is the first of its kind to provide a unified lens for player valuation by extending our current understanding of player evaluation to a market valuation context and integrating players' economic and risk factors, physiological and psychological attributes, and soccer skills in the soccer transfer market research.



Contributing Theories/Fields
Attributes/Features
Target Attribute
Market Efficiency and Equilibrium Conditions
Pricing Theories and Risk Premium
Economic and Risk Factors
Sports Science
Physiological Attributes
Psychological Attributes
Soccer Skills
$\hat{y}$
Player's Market Valuation

Figure 3-1: Conceptual Framework of Soccer Player Valuation

In this framework, the established economic and pricing theories offer some general principles of player valuation. First, efficient valuation should absorb all relevant information. Second, the positive or negative relationship between individual features and market valuation is dictated by the basic law of supply and demand. Nash bargaining theory, as an extension of market equilibrium, explains market valuation in the context of club bargaining power. Third, hedonic pricing theory posits that market value is the sum of the hedonic prices of various individual attributes. Fourth, option pricing theory explains the effect of a few economic and risk factors (e.g., age, injury) from a risk premium standpoint. Sports science is a major contributing field in which more value drivers can be identified. Player valuation can be more efficient by exploring a large number of sports science factors. This research explores player's physiological and psychological attributes and soccer skills in the context of market valuation.

The two blocks of contributing theories and fields in Figure 3-1 allow the exploration of a large number of features in predictive analytics settings. Those features or attributes generally fall into two broad categories: market value creation and market value destruction (Giuliani, 2012). According to Giuliani (2012), a value driver is a variable creating market value, and a negative value driver is a variable diminishing it. Given a target attribute (i.e., a player's market valuation in this research), the positive or negative effect of a feature on the target is the reference for defining that variable as a value driver or a negative value driver (Serna Rodríguez et al., 2019). Within economic and risk factors, certain nationalities, leagues, superstar status, or more broadly, high popularity, could be value drivers. Team strength might be a value driver, as a competent team may increase market value of its players. On the other hand, the number of remaining years in a player's contract are a negative value driver, as bargaining power is conditioned on the remaining years in the contract. Within physiological and psychological attributes and soccer skills, speed, endurance, shooting, passing, dribbling, ball control, and off-ball movement, could be value drivers, to name a few (Herm et al., 2014; Müller et al., 2017). Like some economic and risk factors (e.g., popularity), many of those attributes and skills are rare traits. In the transfer market, players who manifest those traits can be thought of as scarce human capital that are not readily available, thereby receiving a higher valuation. Sports science knowledge and expertise in soccer are needed to explicate why a certain trait is a value driver.

Attributes or features may have an inherent hierarchy. From the bottom up, most physiological and psychological attributes and soccer skills are internal attributes at individual player level. Some economic and risk factors are well above individual level. For example, at least part of popularity stems from external, subjective judgement (e.g., media hype). Bargaining power between clubs depends on buying and selling club characteristics such as their spending power (Tunaru et al., 2005; Franks et al., 2016). In a "money" league (e.g., EPL), financially well-endowed clubs tend to spend huge on the transfer market, while "farm" leagues (e.g., Portuguese Primeira Liga) clubs amass profits by selling their players (Matesanz et al., 2018). Players with high market valuation have flocked into a money league. At a global market level, inflation level and regulation changes affect how a player is valued. The Bosman ruling makes the remaining years in the contract a likely value driver. Player valuation is determined by a mix of intrinsic attributes (e.g., physiological and psychological attributes, soccer skills) and extrinsic, contextual factors like team strength and league (Buekers et al., 2015).

It is also important to note that some attributes or features are not independent of each other. For example, the lack of endurance may be responsible for getting injured more easily in a challenge. As players age, their risk of getting injured usually increases. Some physiological and psychological attributes could be partially dependent on age in specific contexts. Therefore, market value is not only determined by individual attributes that fit into the proposed framework, but also by the interplay between these attributes, as well as the dynamics of environment (e.g., team strength, league). The bundle of physiological and psychological attributes and soccer skills need to materialize on the field through their interactions. For example, fast players could feel sluggish and less responsive, if they lack agility and balance. Similarly, players with high agility, balance and ball control ability will be more likely to attempt a successful dribble.

Accurate player valuation is far from trivial. Little is known about how value drivers influence valuation in a quantitative way. Building upon economic and pricing theories as well as sports science knowledge, this research aims to uncover the key drivers of player valuation not only through theoretical scanning guided by the proposed framework, but also through the application of advanced modeling techniques to handle the complex interaction of various attributes and recognize covert patterns in their relationships with market value. To this end, this research exploits open-source data to expand the sources of potential value drivers and employs predictive analytics as the primary data analysis methodology.

Chapter 4 Data and Methodology

Figure 4-1 depicts the three major phases of the methodology to operationalize the proposed conceptual framework of soccer player valuation. Prior to phase 1, Section 4.1 gives a high-level overview of the data sources. Phase 1 (Section 4.2) describes data pre-processing and feature engineering with a detailed introduction of what attributes are selected. Phase 2 (Section 4.3, Section 4.4) explains the relevant methodology and algorithms that fit the context and objectives of this work and sets up the performance metrics, cross-validation and model tuning. Phase 3 (Section 4.5) introduces some methods for model explanation.

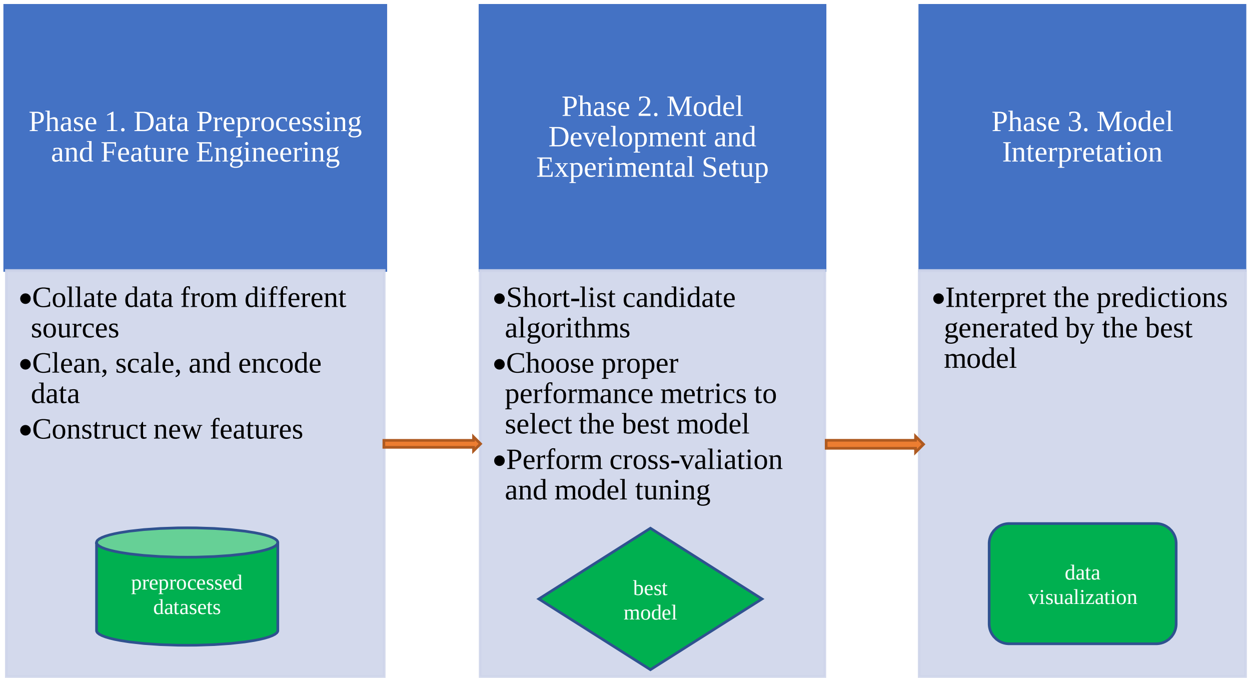

Phase 1. Data Preprocessing and Feature Engineering
•Collate data from different sources
•Clean, scale, and encode data
•Construct new features
preprocessed datasets
Phase 2. Model Development and Experimental Setup
•Short-list candidate algorithms
•Choose proper performance metrics to select the best model
•Perform cross-valiation and model tuning
best model
Phase 3. Model Interpretation
•Interpret the predictions generated by the best model
data visualization

Figure 4-1: Methodology Process

4.1 Data Understanding

In the sports transfer market, researchers have access to detailed data from various sources including player demographics, individual technical skills, physiological, psychological and performance metrics. A preliminary probe of the related work has discovered two highly cited soccer data sources that generate the sheer volume of open-source, publicly available data with reasonable veracity (Herm et al., 2014; Müller et al., 2017; Payyappalli & Zhuang, 2019; Patnaik et al., 2019; Pariath et al., 2018): 1) Sofifa,[4] and 2) Transfermarkt.com.[5] Launched in Germany in 2001, Transfermarkt.com is the leading crowdsourcing website, specializing in gauging market value of soccer players. Beautiful Soup, a Python library for web-scraping, is used in my study to source market value data and club transfer balance (income minus expenditure) from Transfermarkt.com. Due to the difficulties in measuring soccer skills within a laboratory context, there has been limited experimental research carried out in this area (Hoare & Warr, 2000). An alternative statistical procedure is to construct a power score or index that is a proxy for these

[4] https://sofifa.com/
[5] https://www.transfermarkt.us/

fundamental characteristics or explore the latent skills and strengths of the teams (Stekler et al., 2010). Following a similar procedure, Sofifa aggregates rating data from the most popular football simulation video game FIFA that is based on real-world player characteristics and performance. Kaggle, an online community of data scientists, makes the players data scraped from Sofifa.com available for download.

The lack of ground truth in the transfer market research makes it hard to validate and test a model, since a more accurate evaluation of the unobservable, "true" value of a player can be obtained only in retrospect. Every player has a theoretical market value, but not all players are moving from one club to another during transfer seasons. If transfer fee is chosen as the target variable, this type of economic valuation will be unavailable for most players. Moreover, transfer fees sometimes are fraught with inflation, manipulation, speculation and information asymmetry. A consequence is the divergence between the transfer fees paid and the "real" player value, a phenomenon called "hyperinflation" (Magee, 2017). In light of the dearth of an unbiased, optimal player value, data at Transfermarkt.com serves as a reliable benchmark for predicting player value at large scale. Transfermarkt.com is an application of "wisdom of crowds" that has gathered momentum in aggregating individual predictions into realistic estimates of fundamental value in recent years (Surowiecki, 2005). It implements a hierarchical but ad hoc approach called, judge principle, to determine market values such that a few merited community members, also known as "judges", can calculate the final market value by filtering, weighting, and aggregating different opinions the crowd has offered (Herm et al., 2014). The judges have the discretion to override extreme opinions and give more weights to the opinions from more qualified members. Aggregated opinions are often more accurate, leading to better performance (Bachrach et al., 2012). Transfermarkt.com outperforms purely democratic approaches (e.g., computing the mean or median of all individual market-value estimates) and mitigate the risk of bias (Müller et al., 2017). It has a high economic relevance for real-world transfer and salary negotiations (Herm et al., 2014). Therefore, this thesis chooses the market value estimated by Transfermarkt.com as the target attribute.

4.2 Data Preprocessing and Feature Engineering

Each row in Sofifa and Transfermarkt.com dataset represents a player. However, there is no shared unique identifier (i.e., key) for joining the two datasets. Although both have player name columns, a player's name often uses different conventions such as long name, short name, full name, and non-English name. Therefore, player name as such is not a desirable candidate key given its lack of uniqueness. This work first uses player name combined with date of birth as the composite key, matching approximately 80% players in the transfermarkt.com dataset. Then, record linking technique has been implemented to the rest 20% players. Instead of realizing 100% exact match, record linkage is a probabilistic approach akin to fuzzy matching. In this step, the two datasets are joined based on player name (probabilistic match), date of birth (exact match), and club name (probabilistic match). Like player name, club name has a few variations. Empirical threshold for probabilistic match is set to 0.5 for fuzzy match (Stanojevic & Gyarmati, 2016).[6] Table 4-1 is a snippet of the matched players. In Table 4-1, the first column (TM Player Name) and the second column (FIFA Player Name) contain player names used by Transfermarkt.com and Sofifa, respectively. For each row, the player names in the first two columns do not exactly match, but the probability of both names actually being identical is above

[6] A threshold of 1 means the names are identical, and 0 means they have nothing in common.

the 0.5 threshold according to record linkage. Therefore, the similarity score of the two player names is 1. The last column is the total similarity score of player name, club name and birthday. A total score of 3 means similar player names, same club, and identical birthday of the two matched records, which shows a high degree of confidence that the two records describe the same player. For example, player Mikel Merino in the first column and player Mikel Merino Zazón play for the same club and were born in the same day. Thus, it is safe to conclude that both names refer to the same player. To ensure a high accuracy of record linking, only matched records that receive a total similarity score of 3 are included in the final consolidated dataset.

Table 4-1: Matching Player Name by Record Linking

TM Player Name	**FIFA Player Name**	**Similar Player Name**	**Same Club**	**Identical Birthday**	**Total Similarity Score**
Mikel Merino	Mikel Merino Zazón	1	1	1	3
Joan Jordán	Joan Jordán Moreno	1	1	1	3
Sergi Gómez	Sergi Gómez Solà	1	1	1	3
Rúben Vezo	Rúben Miguel Nunes Vezo	1	1	1	3

Feature engineering is the act of extracting numeric quantities from raw data and transforming them into formats that are suitable for the machine learning model and task (Zheng & Casari, 2018, p. 3). Louzada et al. (2016) carried out principal component analysis (PCA) to construct physical, technical and general score of players. Nsolo et al. (n.d.) employed filter method and wrapper method for feature selection. Most filter methods explore the intrinsic properties of the features via univariate statistics instead of cross-validation prior to training the models. Thus, they evaluate each feature in isolation and the scoring of features is independent of the models (Kuhn & Johnson, 2013, p. 499). When dealing with high-dimensional data, it is faster and less computationally cheaper to use filter methods than wrapper methods. Information gain is one of the most widely used filter methods. It calculates the reduction in entropy from the transformation of a dataset. According to information theory, entropy quantifies the average level of information or uncertainty inherent in the value of a random variable or the outcome of a random process (Shannon, 1948). Entropy's significance in the decision-tree-based algorithms is that it provides a means to estimating the heterogeneity of the target variable (Sethneha, 2020). The relationship between the probability and the heterogeneity is expressed in the mathematical form with Equation 1:

$$\boldsymbol{Entropy\ (p)\ =\ -\sum_{i=1}^{N}(p_i \times log_2\, p_i)} \qquad \text{Equation 1}$$

The uncertainty is represented as the log of the probability of a category (p_i) in base 2 where *i* refers to the number of possible categories. The simplest scenario is a binary classification (*i* = 2).

Table 4-2: Feature Construction

Existing Feature	Created Feature	Description	Rationale
age	age group	Labels: youth stage (≤ 23); growth stage (≥24 and ≤28); decline stage (≥28)	The inspiration of age discretization (also known as binning) comes from Majewski and Majewska (2017) and the empirical age distribution in the dataset.
player traits	injury risk	Labels: low, medium, high	The player traits feature (text) describes some special traits that are not manifested in many players and may not be well captured by the numerical features. Players with an "Injury Free" trait have low probability of being injured. Players with an "Injury Prone" trait have high injury probability. Players without both traits have medium injury risk.
player positions	position category	Labels: attacker, midfielder, defender, goalkeeper, substitution, reserve	The player positions feature has 29 labels (specific positions) such as central forward, left midfielder, and right back. For simplicity, the position category feature uses 6 general labels. Substitution players (also known as rotation players or bench players) are part of the first team but are not frequent starters. Reserve players are backup players such as young prospects who have rather limited playing opportunities.
nationality	nation group	Label: France, Italy, England, Germany, Brazil, Argentina, Belgium, Spain, Netherland, Portugal, other countries	The nationality feature has 87 labels (countries). Such a large categorical variable makes one-hot coding less memory efficient. All countries are regrouped into 10 labels.
player overall rating	team rating	The average overall (numerical) rating of all players in each club	To test whether team strength (i.e., how competitive and rich a team is) is a value driver. Team rating may also be an indicator of team bargaining power.
contract valid until	contract remaining	The contract year minus the start year of a given season	To test whether the number of years remaining in the contract is a value driver.

Table 4-3: Feature Selection

Group	Feature	Data Type	Short Description
Economic and Risk Factor	age	numerical	The age of a player in a given season.
	age group	categorical	The discretization for age.
	injury risk	categorical	The chance of a player being injured.
	team rating	numerical	The average overall rating of all players in each club.
	nation group	categorical	The nationality of a player represents.
	league	categorical	The league a player's club belongs to.
	position category	categorical	The general position category or squad status of a player.
	contract remaining	numerical	The number of remaining year(s) in each player’s contract.
	international reputation	numerical	The higher the rating the more famous the player is.
Physiological Attribute	preferred foot	categorical	A player’s preferred/dominant foot, either left foot or right foot, which he uses more frequently and adeptly.
	acceleration	numerical	The higher the rating, the shorter the time needed to reach the maximum sprint speed.
	spring speed	numerical	The higher the rating, the faster the player runs while in full speed.
	agility	numerical	The higher the rating, the more agile the player is while moving or turning.
	reactions	numerical	The higher the rating, the more quickly the player is responding to a situation around him.
	balance	numerical	The higher the rating, the more easily the player is able to maintain balance when facing physical challenges.
	stamina	numerical	High stamina rating means longer time the player can spend sprinting during a game as well as shorter recovery time. Stamina is also responsible for your player getting injured more easily in any challenge
	jumping	numerical	The higher the rating, the higher the player can jump to win aerial balls.
	strength	numerical	The higher the rating the more physically strong the player is.
Psychological Attribute	aggression	numerical	The higher the rating, the more successful tackles and more fouls a player is to commit.
	composure	numerical	The higher the rating, the better the players perform under pressure.
	vision	numerical	The higher the rating, the greater the player’s awareness of the position of his teammates and opponents is. Therefore, the player is more likely to deliver accurate and intricate through balls to set up big score opportunities.
	positioning	numerical	The higher the rating, the more likely a player is to occupy advantageous positions for receiving the ball and attacking the opponent’s goal.

Soccer Technical Skill	finishing	numerical	The higher the rating, the more easily the player shoots on target.
	long shots	numerical	The higher the rating, the more accurate shots from outside the box are.
	shot power	numerical	The higher the rating, the harder the player hit the ball while still keeping a shot accurate.
	penalties	numerical	High penalties rating means the player is good at taking penalties.
	heading accuracy	numerical	The higher the rating, the more accurate a headed pass or header at goal is going to be.
	volleys	numerical	High volley rating means accurate shots taken while the ball is in air.
	free kick accuracy	numerical	The higher the rating the better the accuracy of a direct free kick on goal.
	short passing	numerical	The higher the rating, the faster and more accurate the short or ground pass will be.
	long passing	numerical	The higher the rating, the faster and more accurate the long pass in the air will be.
	dribbling	numerical	A high dribbling rating means the player will be able to keep better possession of the ball whilst dribbling.
	curve	numerical	The higher the rating the more curl the player is capable of putting on the ball when passing and shooting.
	crossing	numerical	High crossing rating means high probability of a medium or long-range pass from a wide area of the field towards the centre of the opponent's box finding the teammate and circumventing the opponents.
	ball control	numerical	The higher the rating, the less likely the ball is to bounce away from the player after receiving it.
	standing tackle	numerical	The higher the rating, the more likely the player is to perform a standing tackle without committing a foul.
	sliding tackle	numerical	The higher the rating, the more likely the player is to perform a sliding tackle without committing a foul.
	marking	numerical	The higher the rating, the more easily the player can track and defend an opposing player.
	interceptions	numerical	The higher the rating, the more likely the player is to catch the opposing team's passes.
	weak foot	numerical	Weak foot is defined as the player's foot other than his preferred foot. High weak foot rating means higher shot power and better ball control for the weak foot of that player.
	gk_kicking	numerical	Goalkeeper's ability to distribute long and accurate goal kicks, from out of the hands or on the ground.
	gk_positioning	numerical	Goalkeeper's ability to position himself correctly when saving shots or reacting to crosses.
	gk_reflexes	numerical	Goalkeeper's agility when making a save.
	gk_diving	numerical	Goalkeeper's ability to make a save whilst diving through the air.
	gk_handling	numerical	Goalkeeper's ability to catch the ball and hold onto it.

Most features in Sofifa are numerical variables on a normalized scale of 1 to 100, while a few are numerical variables on a scale of 1 to 5 (e.g., international reputation). Models that entail matrix manipulation (e.g., linear regression, PCA) are sensitive to the scale of the numerical features (Zheng & Casari, 2018, p. 29). For example, international reputation varies less than balance because of their respective scales (5 vs. 100), PCA might determine the direction of maximal variance more closely corresponds with balance, provided that the two features are not scaled. A unit change in balance can be considered more important than that in international reputation, which is probably incorrect. The most common scaling techniques are standard scaling (variance scaling), min-max scaling, and robust scaling. These manipulations coerce the numerical features to have a common standard deviation, therefore improving the numerical stability of some computations. The trade-off of scaling is a loss of interpretability of the individual features being transformed since those features are no longer in the original units (Kuhn & Johnson, 2013, p. 31). Robust scaling uses the median and quartiles, instead of mean and variance, which ignores outliners such as measurement errors that are very different from the rest (Müller & Guido, 2016, p. 133). The most common technique to represent categorical variables are the one-hot-encoding, also known as dummy variables. The idea is to replace a categorical variable with one or more new features that can have the values 0 and 1 (Müller & Guido, 2016, p. 213).

This work constructs six new features by transforming some existing features. Table 4-2 lays out the existing feature name, the created feature name, a short description, and a rationale of creating the new feature. Table 4-3 presents the final feature set (45 features in total, 39 numerical features, 6 categorical variables) which lies in the intersection of Sofifa and Transfermarkt.com dataset. Features are grouped by economic and risk factor and sports science measurement (i.e., physiological and psychological attributes and soccer skills). The feature set includes new features to complement the value drivers discussed in Chapter 2. Different metrics (e.g., weight, gain) will be used to filter the most important features, which also formalizes the heuristics behind the “wisdom of the crowds” approach Transfermarkt.com takes. There is no universal dataset covering all relevant features. Club bargaining power can be operationalized by aggregating the overall rating of each player in a club.

4.3 Model Development – Machine Learning and Predictive Analytics

Prior research has explored several methods for market value predictions. Payyappalli and Zhuang (2019) used a simple moving average method to forecast market values. This is a very naive approach since the forecasted values were merely based on historical market value data. Majewski & Majewska (2017) approximated the hypothetical value of soccer players’ performance rights via Monte Carlo Simulations. The core idea of Monte-Carlo simulations is to repeatedly use random samples of a well-defined set of parameters or inputs to estimate the mean value of the variable of interest (Martin, 2016). Monte Carlo simulations not only calculate market values but also the distribution of probable market values. However, they only studied four high-profile soccer players, which makes the results much less representative. Empirical modeling in the transfer market literature has been dominated by explanatory statistical modeling where underlying (causal) hypotheses are tested, such as finding determinants of market values through multiple linear regression (Herm et al., 2014). However, explanatory statistical models built for testing hypotheses face the risk of underfitting when it comes to generating empirical predictions (Shmueli & Koppius, 2011). Explanatory models first assume that the data is generated by a certain form of the function (e.g., logistic, exponential, normal) and then find the parameters that give the "best fit" between the data and the function. Linear regression is an example of parametric methods that make an explicit assumption with respect to the functional form of f and have a fixed number of parameters. Although parametric models are fast to deploy, they need more rigid assumptions about the nature of the data distributions. A linear model may vastly deviate from the "true" market value function. Some variables do not have a linear relationship with market value, which violates a fundamental assumption of linear regression. Variables riddled with complex interactions also reduce the accuracy of linear regression. The dominance of explanatory statistical modelling and rarity of predictive powers are a major gap in the transfer market research. This thesis builds upon exploratory modeling for a better understanding of key drivers and extends to predictive analytics approach as its primary methodology for market value prediction. Predictive analytics will empirically validate the conceptual framework at large scale by testing which key drivers can better predict market value.

Machine learning (ML) has been defined by different experts in different contexts. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997, p. 2). Rather than writing explicit and exhaustive instructions step by step, the hallmark of ML is to feed data into algorithms and let the algorithms return patterns automatically. Traditional programming uses data and program as input to generate output (in my case, market value). It is practically infeasible to specify every rule explicitly and manually in program that governs player valuation. Machine learning, by contrast, uses data and market value as input to generate program wherein rules are automatically and implicitly being learned. Predictive analytics is one of the most mature applications of ML. As defined by Kuhn and Johnson (2013, p. 2), predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction. This research attempts to understand the role of the attributes from the various contributing fields and theories in an explanatory phase by developing a prediction model for player market value. This type of tasks is also known as supervised learning since the target attribute (i.e., market value) is known in the dataset. To be clear, player valuation, as one of many business applications of predictive analytics, does not predict or forecast a future outcome *per se*—how much might a player be worth in the next few seasons. On the contrary, this work intends to deliver the current market valuation of a player based on known features. Procedurally, the goal is to apply statistical learning methods to a training data set in order to estimate the unknown function f to compute market values Y based on a vector of player attributes X. In other words, the purpose is to find a function $\widehat{f}$ such that $Y \approx \widehat{f}\,(X)$ for any observation (X, Y). To overcome the limitations of parametric methods, flexible non-parametric models that can fit multiple possible functional forms for f are chosen, because non-parametric methods do not require explicit assumptions as to the particular functional form of f. They are more likely to fit a wider array of candidate forms and obtain an estimate as close to the data as possible without overfitting (James et al., 2017, p. 23). Since non-parametric approaches do not reduce the problem of estimating f to a small number of parameters, to obtain an accurate estimate for f, far more observations are required than what are typically needed for a parametric approach.

While developing a predictive model is a key goal of this work, balancing accuracy and explanation will be the guiding principle in the process. "All models are wrong, but some models are useful" (Box & Draper, 1987, p. 424). In predictive analytics, various forms of models can be devised using different algorithms that can learn predictive patterns from a training data set. The specific choice depends on speed-accuracy-complexity tradeoffs (Murphy, 2012, p. 25). At a high level, any ML algorithm has three main components (Domingos, 2012): 1) representation, 2) evaluation, and 3) optimization. Tree-based and rule-based models have three intrinsic advantages for player valuation. First, these forms of predictive models can be explained such that the decision-making steps can be visualized by a tree representation. Second, they can effectively handle heterogeneous predictors (continuous, categorical, etc.) and predictors with missing data points. In addition, they implicitly conduct feature selection. Third, these models do not require specific assumptions about the form of the predictors' relationship with the target (e.g., linear). Five candidate models are short-listed for this work, namely, Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR) and Extreme Gradient Boosting (XGBoost). All models are employed in a regression setting.

4.3.1 Multiple Linear Regression (MLR)

The MLR model can be written in Equation 2:

$$\boldsymbol{f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j} \qquad \text{Equation 2}$$

An input vector $X^T = (X_1, X_2, \cdots, X_P)$ to predict a real-valued output Y. β_0 represents the estimated intercept. Here the β_j's are unknown parameters or coefficients (linear in the parameters), and the variables X_j can come from various sources including quantitative inputs, transformations of quantitative inputs (e.g., log), numeric or "dummy" coding of the levels of qualitative inputs, and interactions between variables (Hastie et al., 2009, p. 44). A MLR model seeks to estimate the parameters β so that a function of the sum of the squared errors is minimized. The most popular estimation method is least square. A distinct advantage of MLR models is that they are highly interpretable. The estimated coefficient of a predictor equals to the number of units increase or decrease in the response variable given 1 unit increase in that predictor.

4.3.2 Decision Tree (DT)

As mentioned above, one advantage of DT is its inherent methods or heuristics to choose influential features for prediction (Daumé, 2017, Chapter 2.). In addition to its interpretability, this will help improve model accuracy without increasing complexity. From a programming standpoint, DT can be conceptualized as a cascade of if-then statements. For example: *Is this player older than 26? Does this player have a shooting ability greater than 80?* Regression trees partition the data into smaller regions that are more homogenous with respect to the target variable (Kuhn & Johnson, p. 175). To achieve outcome homogeneity, regression trees automatically search the feature to split on and split value of that feature. Then, the multidimensional feature space—that is, the set of possible values for X_1, X_2, . . ., X_p—is divided into J distinct and non-overlapping regions, R_1, R_2, . . ., R_j. In theory, those regions could have any topology. However, high-dimensional rectangles, or boxes, are desirable for simplicity and interpretation purpose (James et al., 2017, p. 306). For every observation that falls into the region R_j, regression trees make the same prediction $\hat{y}_{R_j}$, which is the average of the training set outcomes within R_j. One of the most utilized techniques for constructing regression trees is the classification and regression tree (CART) methodology of Breiman et al. (1984). Equation 3 is the optimization objective function to find boxes $R_1, \ldots, R_j$ such that the overall sums of squares error are minimized:

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \qquad \text{Equation 3}$$

4.3.3 Random Forest (RF)

Models based on single trees or rules, however, have two vulnerabilities: (1) model instability (i.e., slight changes in the training data can drastically alter the structure of the tree or rules and, hence, the interpretation) and (2) less-than-optimal predictive performance. To overcome these problems, experts developed ensemble methods that combine many trees (or rule-based models) into one model. The intuition behind ensemble learning is to learn multiple models instead of one model. Ensembles tend to have much better predictive performance than single trees (Kuhn & Johnson, 2013, p. 198). RF, as its name implies, comprises a large number of relatively uncorrelated individual decision trees operating as an ensemble (committee) that will outperform any of the individual constituent tree. Each tree in the RF predicts a market value and the prediction with most votes becomes the final prediction (Stanojevic & Gyarmati, 2016). In some sense, RF imitates the wisdom of crowds Transfermarkt.com embraces (Yiu, 2019).

4.3.4 Support Vector Regression (SVR)

The no-free-lunch theorem proves that there is no universally best model due to the fact that a set of assumptions that work well in one domain may work poorly in another (Wolpert, 1996). Researchers should compare different models in terms of their prediction performance and choose the one that best fits the data. Unlike rule-based learning that utilizes explicit generalization, instance-based learning is a family of highly flexible algorithms that compares a new problem instance with instances seen in the training phase. SVR is an instance-based learning algorithm. A subset of the training examples X is referred to as the support vectors necessary for determining the decision boundaries. SVR is formally defined by an optimal hyperplane that has a maximum number of training data points within the decision boundary. The main aim is to search a decision boundary at epsilon distance from the optimal hyperplane such that data points close to the hyperplane or the support vectors X are within that decision boundary. SVR has a metric called the margin that roughly equals to the distance between the decision boundary and the closest training data point. The margin serves as a buffer that tolerates the error (ϵ_i) below a certain threshold and lest the model makes many rigid assumptions. The slope and intercept of the decision boundary that maximize the margin between the boundary and the data is known as the support vector machines (Kuhn & Johnson, 2013, p. 344). Their weights are represented by a vector α_i. Equation 4 is a non-linear kernel function $K(\cdot)$ that returns the similarity between a new example x_i and those supporting examples X, which corresponds to a dot product (i.e., $x_i' u$). The new examples enter into the prediction function as sum of dot products with the new sample values. When examples enter the model linearly, the kernel function reduces to a simple sum of cross products. With the bias term β_0, the constructed SVR model is given by Equation 5:

$$\boldsymbol{K(x_i, u) = \sum_{j=1}^{P} x_{ij}\, u_j = x_i' u} \qquad \text{Equation 4}$$

$$\boldsymbol{f(u) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, u)} \qquad \text{Equation 5}$$

To estimate the model parameters, SVR uses the ϵ loss function but adds a penalty. The SVM regression coefficients minimize Equation 6:

$$\boldsymbol{C \sum_{i=1}^{n} L_\epsilon(yi - \hat{y}_i) + \sum_{j=1}^{P} \beta_j^2} \qquad \text{Equation 6}$$

$L_\epsilon(\cdot)$ is the ϵ-insensitive function, given a threshold set by the user (denoted as ϵ). The C parameter is the cost penalty that is set by the user, which penalizes large residuals. Data points with residuals less than ϵ do not contribute to the regression fit while data points with an absolute difference greater ϵ contribute a linear-scale amount (Kuhn & Johnson, p. 153).

4.3.5 Extreme Gradient Boosting (XGBoost)

Boosting is another ensemble approach for improving the predictions resulting from a decision tree. Boosting can be applied to many statistical learning methods for regression or classification (James et al., 2017, pp. 321-322). Tree boosting normally begins with a weak learner (e.g., regression trees) and then finds an additive model that minimizes a loss function such as squared error for regression (Kuhn & Johnson, 2013, p. 204). As a member of the family of gradient boosting, XGBoost expects to have the weak learners which perform poorly. When all predictions are combined, poor ones cancel out and better ones form final predictions. As with all supervised learning models, to learn the trees is to define an objective function and optimize that function. In general, the objective function contains a loss function $L(\theta)$ and a regularization term $\Omega(\theta)$. The most common training loss function in XGBoost for regression problems is mean squared error. Equation 8 is a weight function where T is the number of leaves. The objective function is given by Equation 7 where t denotes the number of trees:

$$\boldsymbol{\sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(fi)} \qquad \text{Equation 7}$$

$$\boldsymbol{f_t(x) = w_{q(x)}}, \text{ where } \boldsymbol{w \in R^T}, \boldsymbol{q: R^d \to \{1, 2, \cdots, T\}}. \qquad \text{Equation 8}$$

$$\boldsymbol{\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2} \qquad \text{Equation 9}$$

XGBoost is a scalable machine learning system for tree boosting that has achieved high accuracy across a wide range of domain specific problems in major machine learning challenges such as Kaggle competitions (Chen & Guestrin, 2016). Two advantages make XGBoost a promising candidate for player valuation: 1) parallel processing, distributed computing, and out-of-core computation make processing massive datasets faster and more resource efficient; 2) incremental learning enables continual training on the existing model from its last iteration (Aarshay, 2016). The major disadvantage is the relatively low model interpretability, because XGBoost is a black-box algorithm which has complex internal representation.

4.4 Experimental Setup

4.4.1 Performance Metrics and Model Evaluation

It is hard to assert which of the aforementioned algorithms estimate player valuation best without implementing and running all of them on the datasets. Experiment setup specifies which algorithms to run on which datasets, including how these settings are chosen; how the algorithms are evaluated (Blockeel & Vanschoren, 2007). MLR is the baseline model of player valuation, followed by DT, RF, SVT, and XGBoost. As the models get more sophisticated, RMSE and adjusted R^2 may expect to decline and improve, respectively. However, the complex models are more computationally intensive and less explainable. After acquiring the experiment results, the specific model choice depends on speed-accuracy-complexity tradeoffs (Murphy, 2012, p. 25). Table 4-4 lists some metrics that are most commonly used to evaluate and compare individual ML regressor models.

Table 4-4: Common Performance Metrics in ML Regression Models

Metric	Mathematical Formulation	Note
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$	The root of the expected value of the squared error.
Mean Absolute Error (MAE)	$MAE = \frac{\sum_{i=1}^{n} \lvert \hat{y}_i - y_i \rvert}{n}$	The expected value of the absolute error.
R-Squared	$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}$	It provides an indication of goodness of fit.
Adjusted R-Squared	$R_{a\,dj}^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \times \frac{n-1}{n-k-1}$	It penalizes adding independent variables that do not fit the model.

The table presents the metric name, its mathematical formula, and a short note. RMSE (Root Mean Square Error) is an evaluation metrics for numerical target variables (e.g., market value). As it is a squared term, large errors will be amplified and increase RSME and lead to worse model performance. The objective is to minimize RSME. In particular, RMSE penalizes predictions that are significantly off the actual values than predictions that are marginally off. Many business applications use default R^2 or MAE (Mean Absolute Error) as evaluation metrics for regression tasks (Müller & Guido, 2016, p. 299). R^2 is used to determine the amount of variability in the target variable that is explained by the independent variables in the model. In Table 4-4, y_i denotes the target variable and $\hat{y}_i$ is the estimated value of y_i; n is the number of instances used to fit the model and k is the number of features in the model. A high R^2 can be misleading, as it always increases with every independent variable added to a model. Adjusted R^2 is a better estimate of explained variance than default R^2, because the former penalizes adding irrelevant independent variables. This study uses both RMSE and adjusted R^2 for model selection.

4.4.2 Cross-Validation and Model Tuning

Without proper sampling and data partitioning strategies, complex ML models can lead to a phenomenon known as overfitting, which means they follow the noise mechanically and fare poorly when predicting a previously unseen instance (James et al., 2017, p. 22). Cross validation is a common sampling strategy to ameliorate overfitting. K-fold cross-validation randomly partitions the training dataset into *k* distinct subsets (called folds) of roughly equal size. Then, a model is fit using all folds but the first one. The model predicts the value of the first fold and performance measures are evaluated, so on and so forth (train and evaluate the model *k* times). Each time, the fold solely for evaluation purpose is also known as held-out samples. At the end, performance measures are summarized usually with the mean and standard error (Kuhn & Johnson, 2013, pp. 69-70). Figure 4-2 is a schematic diagram of 5-fold cross-validation. Cross validation generally involves shuffling the order of the instances and therefore should be cautiously used to split the data into training and testing set for the sports prediction problem (Bunker & Thabtah, 2019). A held-out training test split is recommended, with the time order of the instances being preserved. This is essentially like order-preserved leave-one-out cross validation. For each season from this validation period, the model was trained on all preceding seasons (Hubáček et al., 2019). Table 4-5 details the sampling strategy of this work, including two consecutive seasons' data (2018/2019 and 2019/2020), the data size, the training testing split, the models that run on each dataset, and the main purpose of each season's data. 2018/2019 season is solely used for choosing the best model (i.e., model selection) from MLR, DT, RF, SVR, and XGBoost. Next, 2019/2020 season's data (the most recent data available) is used for fine-tuning and testing the best model in the previous season. All footballers in these two datasets play for the Big 5 leagues (English Premier League, Italian Serie A, Spanish La Liga, German Bundesliga, and France Ligue 1) which accounted for almost 75% of global transfer spending in January 2020 (FIFA, 2020b).

Predictive algorithms work with a range of fine-tuning parameters (a.k.a. hyperparameter) that enable the model to learn the underlying structure in the data without overfitting or underfitting. Hyperparameters cannot be directly estimated from the data because there is no analytical formula available to automatically calculate an appropriate value (Kuhn & Johnson, 2013, pp. 64-65). Hence, the existing data should be used to identify those hyperparameters that yield the best and most realistic predictive performance, a process known as model tuning. It can be achieved by splitting the existing data into training and test subsets. Model tuning is often used in conjunction with cross validation. The training set is used to build and tune the model and the test set is used to estimate the model's predictive performance. Modern approaches to model building split the data into multiple training and testing sets, which have been shown to often find more optimal tuning parameters and yield better predictive performance (Kuhn & Johnson, 2013, pp. 61-62). Model tuning is more of art than science. A rule of thumb is to try out consecutive powers of 10 or a smaller number for a more fine-grained search (Géron, 2017, p. 72). Instead of fiddling with the hyperparameters manually, this work uses grid search to try out possible combinations of the hyperparameters of interest. Grid search specifies which hyperparameters and what values to experiment with, once model performance has been summarized across sets of tuning parameters, the simplest philosophy is to choose the final settings associated with the numerically best performance estimates (Kuhn & Johnson, 2013, p. 74). This work uses Sciki-Learn for ML model implementation and. Sciki-Learn is a free, open-source software machine learning library written in Python. All code is experimented on Google Colab, a web-based IDE (Integrated Development Environment)

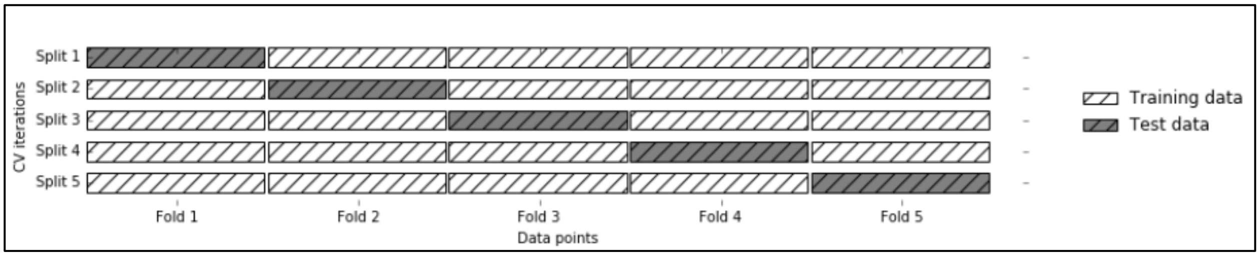


Figure 4-2: Data splitting in 5-fold validation (Müller & Guido, 2016, p. 252)

Table 4-5: Model Training and Test Options

Dataset	**Total Players *N***	**Model**	**Training, CV, and Tuning**	**Test**	**Purpose**
2018-2019 season	2025	MLR, DT, RF, SVR, XGBoost	80%	20%	Model Selection
2019-2020 season	2266	The best model	60%	40%	Interpretation

4.5 Model Interpretability

Understanding how a model makes predictions is crucial for trust, actionability, accountability, debugging, and transparency (Lundberg et al., 2019). For instance, which features have substantial influence on the target variable. To do so, several types of importance values estimating a feature's true effect on the model's output can be attributed to each input feature, including gain and weight. An important distinction is feature importance and feature effect. The former often rely on some processes (e.g., backwards elimination and forward selection in multiple regression) or variance-based measures, whereas the latter expresses how a change in a feature changes the predicted outcome such as partial dependence plots (Molnar et al., 2020). However, those common feature attribution methods for tree ensembles sometimes produce inconsistent results such that a feature with a larger attribution value might actually be less important than another feature with a smaller attribution value, which hinders reliable comparison of attribution values across features (Lundberg et al., 2019).

Brownlee (2016) has a plain explanation on feature Importance of XGBoost: feature importance is a numerical score that indicates how relevant each feature is in the construction of the boosted decision trees. This importance is available for each feature in the final dataset and features are ranked based on their relative importance. Weight, cover, and gain are three metrics of feature importance. The weight, also known as frequency or split count, refers to the number of times a feature is used to split the data across all the boosting decision trees. Since feature splits are chosen to be the most informative, this can represent a feature's importance (Lundberg et al., 2019). The cover metric means the relative number of observations related to this feature.

Both weight and cover are less indicative of the predictive contribution of a feature for the model. A classic approach to feature importance is based on the gain metric. Gain is the total reduction of loss or impurity contributed by all splits for a given feature (Lundberg et al., 2019). The gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. The gain is the most relevant attribute to interpret the relative importance of each feature. A higher value of this metric when compared to another feature implies it is more important for generating a prediction.

Interpretability is a key factor when a ML model is deployed for high-stakes decision-making like player valuation. Interpretable machine learning (IML) methods can be used to discover hidden knowledge, to justify the model and its predictions, and to further improve the model (Molnar et al., 2020). IML methods often create a second (*post hoc*) model to explain the first black-box model (Rudin, 2018). Recently, a novel IML technique known as Shapley Additive Explanations (SHAP) has demonstrated its effectiveness in explaining various supervised learning models (Antwarg et al., 2020). The SHAP values apply a game theoretic framework to ML models where features (i.e., the players) collaborate to make a prediction (i.e., the payout), and generate an explanation as to how a prediction is affected by features in the context of a collaborative game (Molnar et al., 2020). SHAP values have an explanation model g that is a linear function of simplified binary variables:

$$\boldsymbol{g(z') = \Phi_0 + \sum_{i=1}^{M} \phi i z'_i} \qquad \text{Equation 10}$$

M is the number of input features, and $\phi i \in \mathbb{R}$. The z'_i typically represent a feature being observed (z'_i= 1) or absent (z'_i= 0), and the ϕi's are the feature attribution values. SHAP explains a specific prediction by assigning an importance value (SHAP value) to each feature:

$$\boldsymbol{\phi_i = \sum_{s \subseteq N\{i\}} \frac{|s|!(n-|s|-1)!}{n!} [v(S \cup \{i\}) - v(S)]} \qquad \text{Equation 11}$$

Group N is a set of all n input features. S is the set of non-zero indices in z'_i. In SHAP, the contribution of each feature on the model output $v(N)$ is allocated based on their marginal contribution (Shapley, 1953). $v(S)$ is the expected value of the function conditioned on a subset S of the input features. SHAP values combine these conditional expectations with the classic Shapley values from game theory to attribute ϕi to each feature (Lundberg et al., 2019). SHAP values have sound theoretical basis in game theory and consistently attribute feature importance, better align with human intuition, and better recover influential features (Lundberg et al., 2019).

Chapter 5 Analysis and Results

Guided by the market efficiency theory, Chapter 4 compiles an extensive list of features in which value drivers could be identified and puts forward an entire data methodology. This chapter presents the findings of data analysis. Specifically, section 5.1 visualizes the distribution of the target variable and some variables of economic relevance, the correlations between all numerical features, and 20 features with most information gain. Section 5.2 documents the experiment results of machine learning, including the best model, its set of optimal tuning parameters and performance on the testing data. Section 5.3 uses SHAP values to decompose two individual predictions as a small case study and interpret the collective behavior of the best model, which approaches the overarching research question in a novel way.

5.1 Exploratory Data Analysis (EDA)

EDA is an approach to summarizing the main characteristics of datasets, often using visualization methods. The exploratory phase can help better understand the following experiment results and model interpretation. It is very helpful to examine feature and target distribution before implementing any ML models, especially when some variables have uneven distributions. The market value distribution (see Figure 5-1) is right-skewed due to the very limited number of the most valuable players. The highest market value is €200M and the lowest value is barely €75,000. The average value is €10.27M with a very large standard deviation of €18.53M. Models that are trained by this unbalanced distribution will be likely to produce high bias when it comes to valuable players, because the models have very few samples of such players to learn. To resolve skewness, a log transformation is performed so that the log values (See Figure 5-2) are more normally distributed.

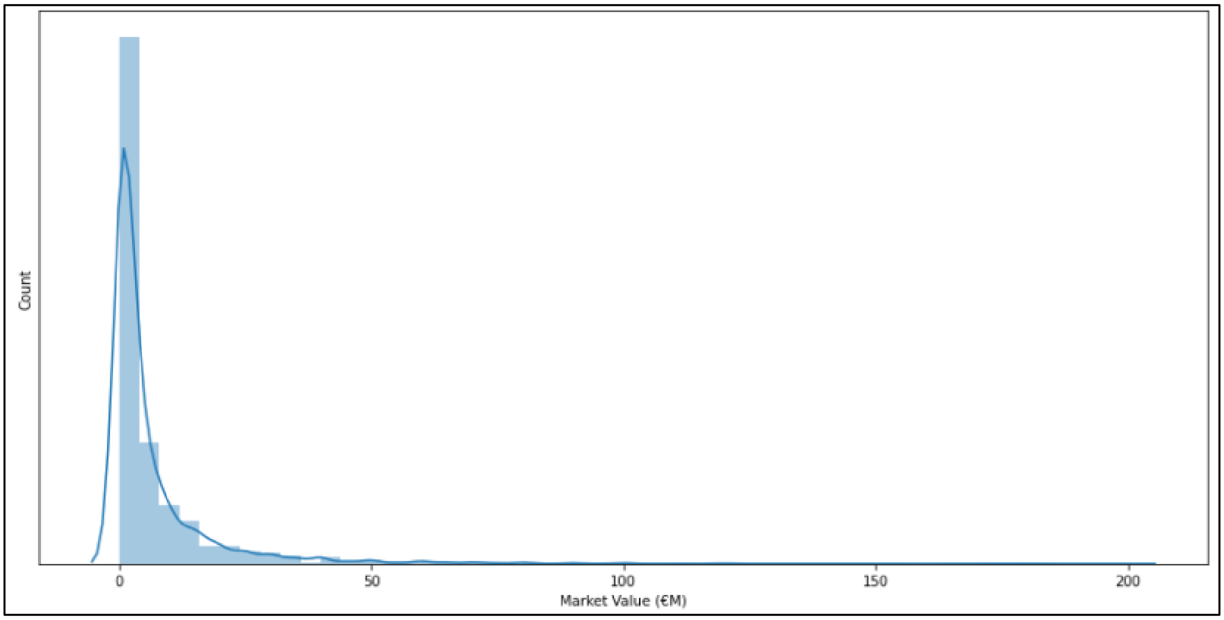


Figure 5-1: Distribution of Market Value (€M)

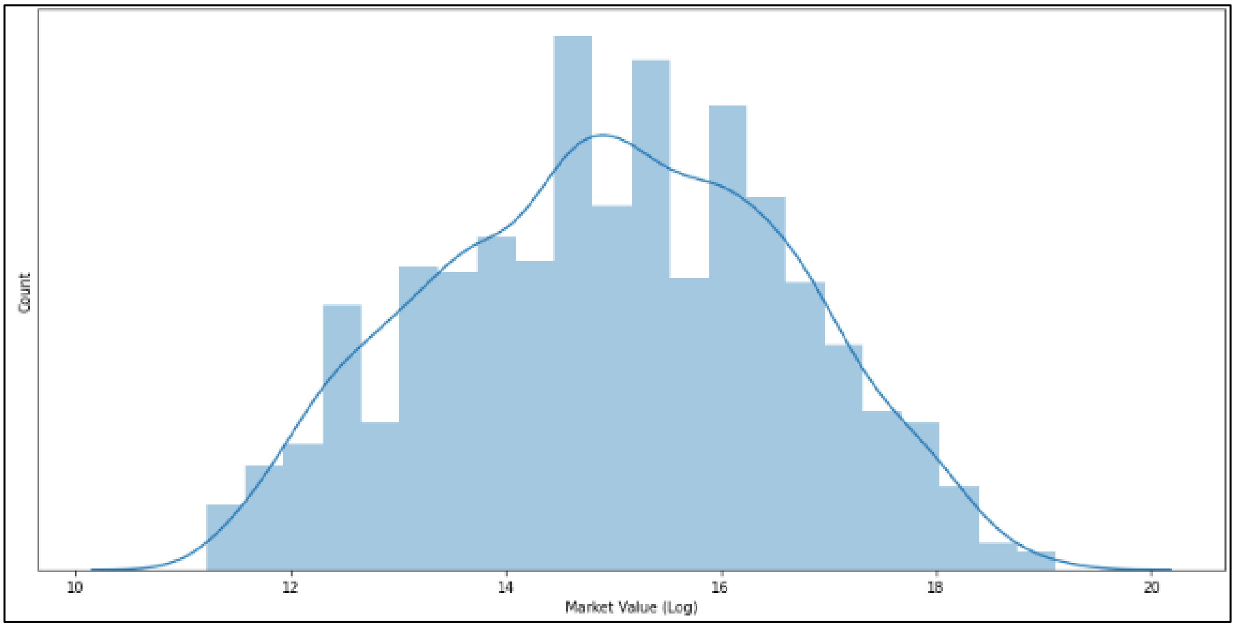


Figure 5-2: Distribution of Market Value (Log)

Age has a visible right-skewed distribution (See Figure 5-3), meaning there are more young players than older players. The average age is 25.2 with a median of 25. International reputation is substantially right skewed (See Figure 5-4). The majority of players are little known (i.e., international reputation = 1), while only a handful are well-known (international reputation = 4) or household names (international reputation = 5) such as Messi and Cristiano Ronaldo.

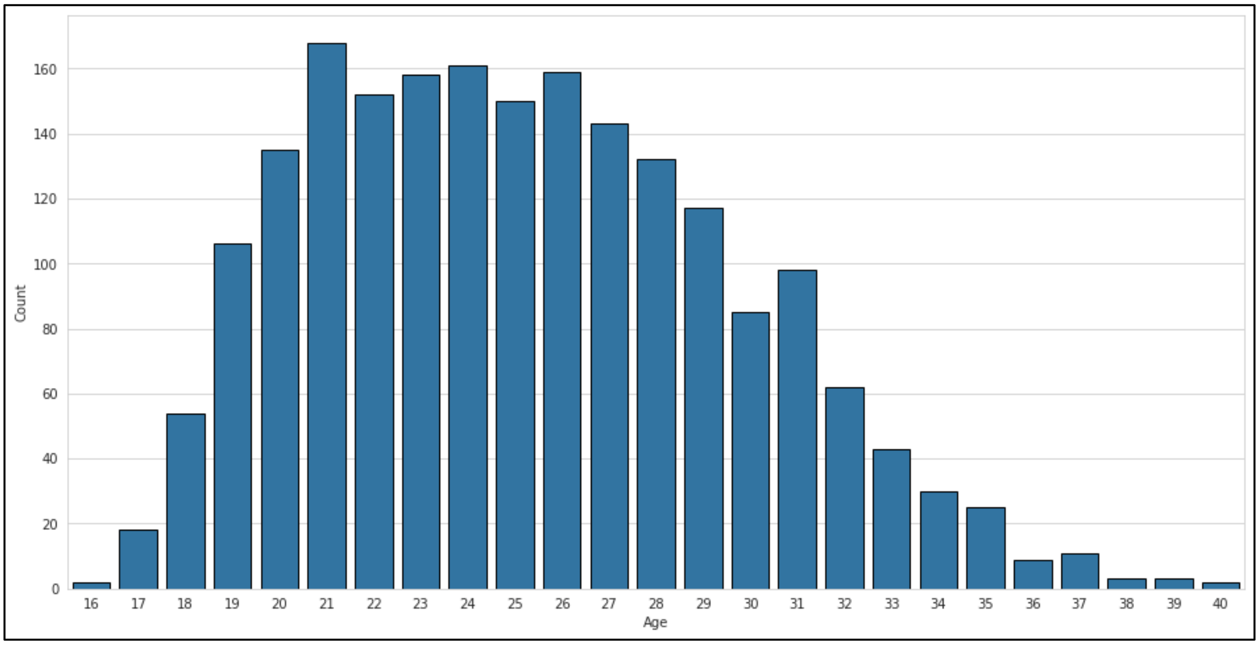


Figure 5-3: Player Age Distribution

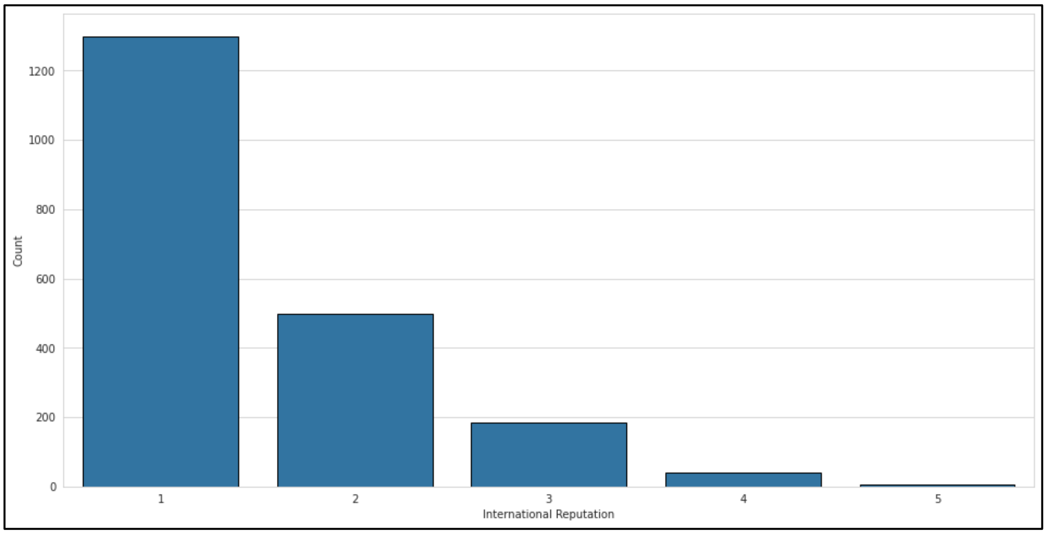


Figure 5-4: Player International Reputation Distribution

Injury risk, position category, and nationality all have imbalanced data representation (akin to imbalanced class). In Figure 5-5, low injury risk and high injury risk account for very low proportions (about 12% combined) in the training data as compared to medium injury risk (88%), which might make this feature less informative. This work is interested in testing whether low injury risk is a value driver and high injury risk is negative value driver. As explained in Chapter 4, player position, a categorical feature with high cardinality is converted into more general positions. In Figure 5-6, nearly half of players (about 43% or 888 out of 2026) are classified as substitution players. This is followed by defender (about 14% or 291 out of 2026), midfielder (about 13% or 269 out of 2026), attacking player (about 9% or 184 out of 2026). In Figure 5-7, nationality has a heavily skewed distribution such that countries have robust youth academy systems like France (e.g., the elite Clairefontaine) are likely to be overrepresented while many small countries being underrepresented (e.g., only 1 player).

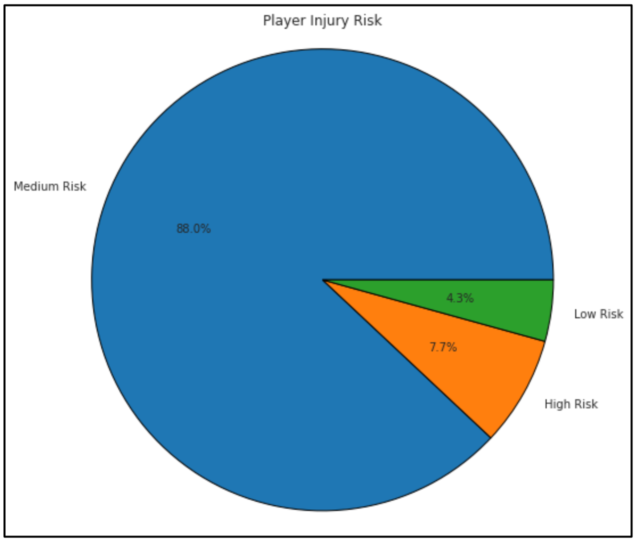


Figure 5-5: Player Injury Risk Distribution

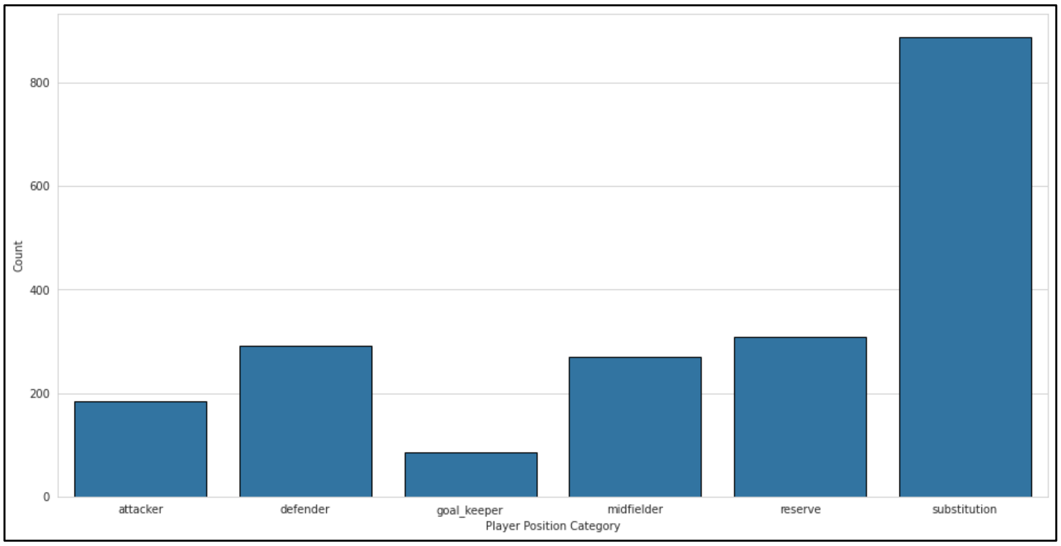


Figure 5-6: Player Position General Category Histogram

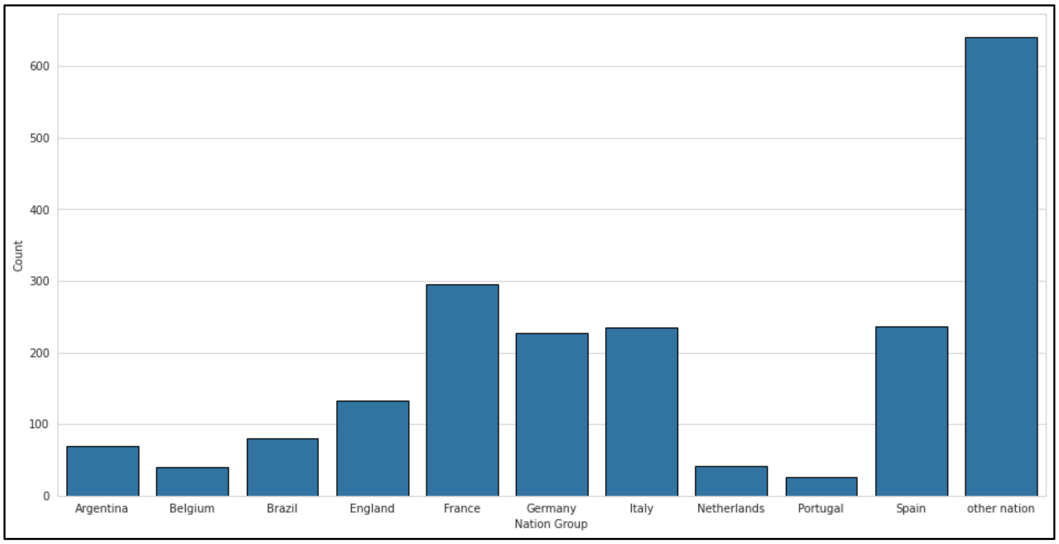


Figure 5-7: Player Nationality Distribution

Multicollinearity is a common problem for causal inference but much less so for association analysis. All models in this work should not be treated as causal models. According to Figure 5-8, a number of features are correlated with each other. Team rating and international reputation are positively correlated. Superstars and renowned players have been flocking into strong clubs (i.e., high team rating). Age and contract remaining have a negative correlation, although it is weak. In practice, clubs have economic incentives to secure long-term contracts of young players while tending to offer short-term contracts to old players. Soccer-specific technical skills under each general skill category are correlated. For example, marking, standing tackle, and sliding tackle are all defensive skills and appear to have very positive correlations. Most notably, the goalkeeping attributes (kicking, positioning, reflexes, diving, and handling) are highly correlated with each other and negatively correlated with nearly all other features. Goal keepers and defenders have a narrow set of standard skills. The attacking skills are not as strongly correlated. However, correlation does not imply causation.

As this works deals with 45 features, exploratory analysis of all the features may be infeasible. Information gain, as a filter method, calculates the reduction in entropy from the transformation of a dataset. Although it can be used for feature selection by evaluating the information gain of each feature in the context of the market value, this work did not select or drop any features before the full modelling step. Figure 5-9 ranks 20 features with strong relationships with the market value might give a hint as to which features are most informative, as these features could be corroborated or be contradicted by model-based feature importance techniques used in section 5.3. Reaction has the largest information gain, followed by composure and team rating. International reputation, age, and contract remaining are among the top 20 features.

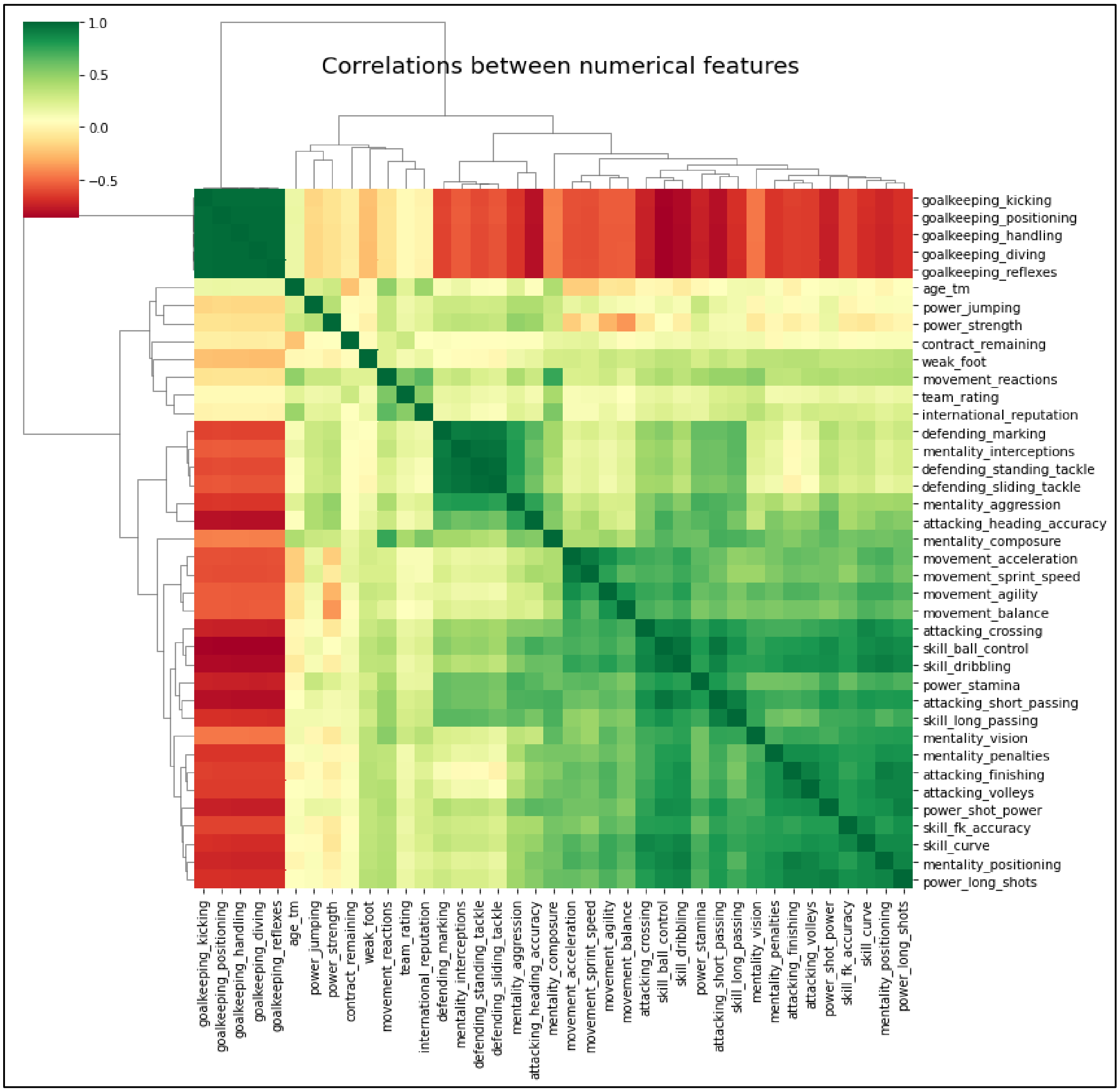


Figure 5-8: Correlation Analysis

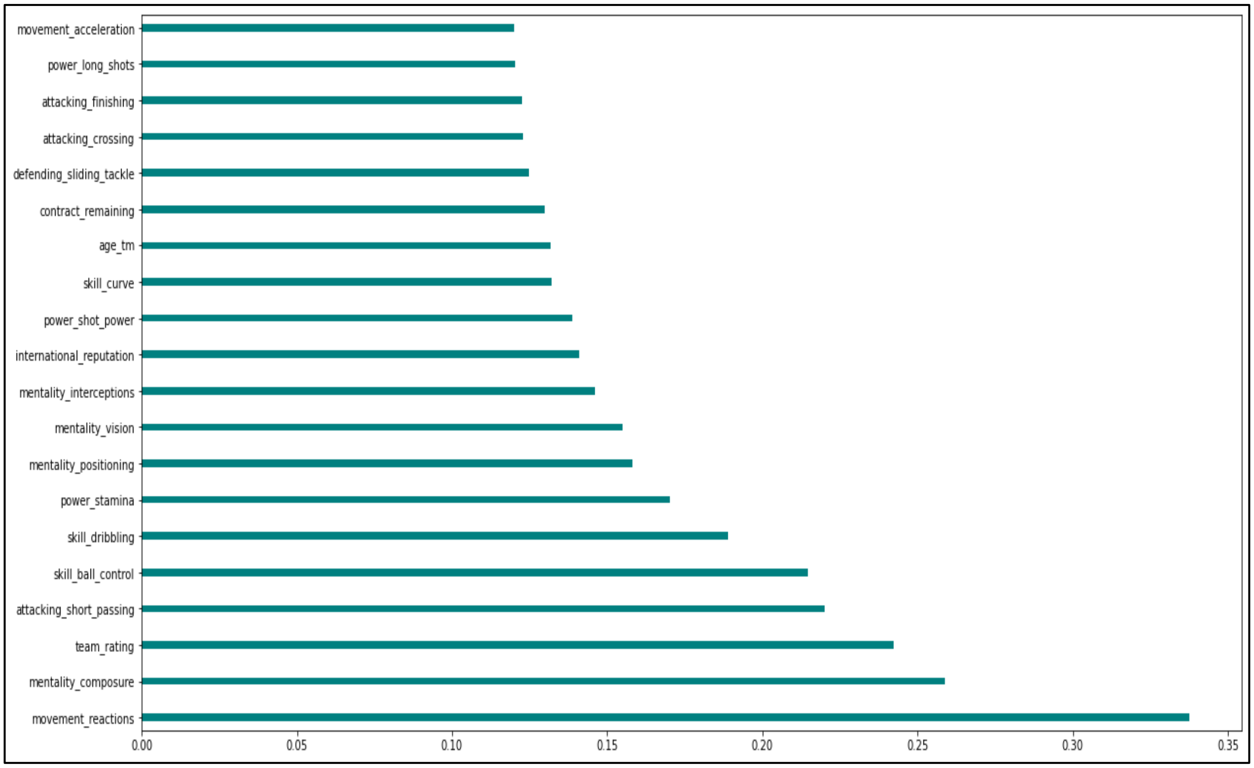


Figure 5-9: 20 Features with Most Information Gain

5.2 Predictive Analytics Performance Results

As noted in section 4.4, 80% of 2018/2019 season's data is for training the five candidate models, while the rest 20% for evaluating their performance. 10-fold cross validation is more computationally intensive. Google Colab has a restriction on free GPU (Graphic Processing Unit) available. In light of this, this study chooses 5-fold cross validation to cope with limited computing resources. As Table 5-1shows, DT has the highest RMSE and lowest adjusted R^2, which is followed by RF and MLR. Surprisingly, MLR outperforms DT and is on a par with RF. Parsimonious models are not theoretically inferior to sophisticated models. The most straightforward way is picking the model that yields the least RMSE and the highest adjusted R^2. In this work, XGBoost is the model with best performance. Although SVR yields close results, it has a significantly longer execution time, which could be a disadvantage from a model deployment viewpoint.

Table 5-1: Model Selection Results

Model	RMSE	Adjusted R^2
Multiple Linear Regression	0.841	0.661
Decision Tree	0.969	0.573
XGBoost	0.721	0.763
Random Forest	0.843	0.676
SVR	0.731	0.757

For clarity, the best model in Table 5-1 is named XGBoost Model V1. To further test the generalizability of XGBoost, this work develops XGBoost Model V2 and XGBoost Model V3.

Table 5-2 lists the tuning parameters of XGBoost this work has configured after 5-fold cross-validation, including parameters names, descriptions, and optimal values. Both XGBoost Model V2 and XGBoost Model V3 use the same set of optimal tuning parameters as shown in Table 5-2. The only difference is that V2 is trained on season 2019/2020 while V3 also uses season 2018/2019 as part of the training data. Both V2 and V3 are tested on the rest 40% of season 2019/2020. V3 has lower RMSE and higher adjusted R^2 than V2. More training data is likely to be attributed to the incremental improvement of predictive performance of V3. The model testing results and the datasets the three models build upon can be found in Table 5-3. Although these models should not be equated with causal models, XGBoost is making inferences on previously unseen players in a consistent way. It is safe to say that XGBoost Model V3 should be reasonably good at generalization.

Table 5-2: XGBoost Tuning Parameters

Tuning Parameter	**Description**	**Default Value**	**Optimal Value**
number of trees B	B is also known as the number of estimators. Unlike random forests, XGBoost can overfit if B is too large.	a relatively small number of trees (e.g., 100 trees).	100
learning rate λ	λ is a small positive number that controls the rate at which boosting learns. Unlike fitting a single large decision tree to the data, the boosting approach instead learns slowly (James et al., 2017, p. 322).	Typical values are 0.01 or 0.001.	0.1
max depth	The max depth is the maximum number of nodes allowed from the root to the farthest leaf of a tree. Deeper trees can model more complex relationships by adding more nodes, but sometimes end up following noise, causing the model to overfit.	The default number of the max depth is 6.	3
min child weight	The min child weight is the minimum weight (or number of samples if all samples have a weight of 1) required in order to create a new node in the tree. A smaller min child weight allows the algorithm to create children that correspond to fewer samples, thus allowing for more complex trees, but again, more likely to overfit.	The default number of the min child weight is 1.	7

Table 5-3: Model Testing Results

Model	Training Data	Testing Data	RMSE	Adjusted R^2
XGBoost Model V1	80% of 2018/2019 season	20% of 2018/2019 season	0.721	0.763
XGBoost Model V2	60% of 2019/2020 season	40% of 2019/2020 season	0.740	0.744
XGBoost Model V3	100% of 2018/2019 season and 60% of 2019/2020 season	40% of 2019/2020 season	0.680	0.784

To interpret the best XGBoost model (i.e., XGBoost Model V3), the ensuing step is identifying the most features. According to Figure A-2 in Appendix, reaction is the most importance feature, followed by reserve status, ball control, team rating, international reputation and contract remaining, which constitute the top six important features. Age and substitution status are also among the top features. Although only 4.3% of players have low injury risk trait, the gain metric views this trait as slightly more important than many more soccer-specific attributes like shorting passing, dribbling, and spring speed. Within the 10 nation groups, German nationality is the most important one, despite that France have some most expensive players. England is not a particularly relevant feature. Perhaps most unexpectedly, Brazilian nationality is deemed the least importance feature. As discussed in section 4.5, the gain metric could be inconsistent and does not distinguish a positive or negative effect.

SHAP values interpret the impact of taking a certain value for a given feature in comparison to the prediction a model would make if that feature took some baseline value. For example, to what extent is a market value prediction driven by the fact that the player is 22 years old instead of some baseline number of age (average age)? SHAP values decompose a single individualized (local) a market value prediction to show the impact of each feature with the following equation:

$$\sum_{i=1}^{n} \phi_i^{(j)} = \hat{y}_j - \bar{y}$$

The sum of SHAP value ϕ_i for the feature i of a player j equals predicted market value $\hat{y}_j$ minus predicted base market value $\bar{y}$. That is, the SHAP values of all n features sum up to explain why a prediction was different from the baseline. This allows us to bread down a prediction in a graph like this:

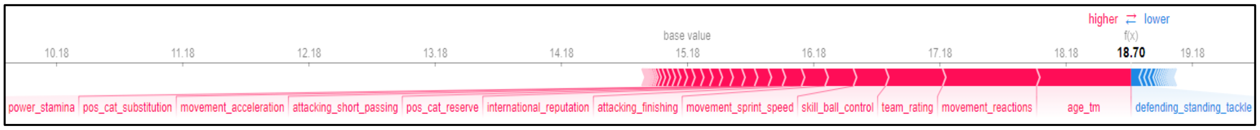


Figure 5-10: Kylian Mbappé Prediction — SHAP values

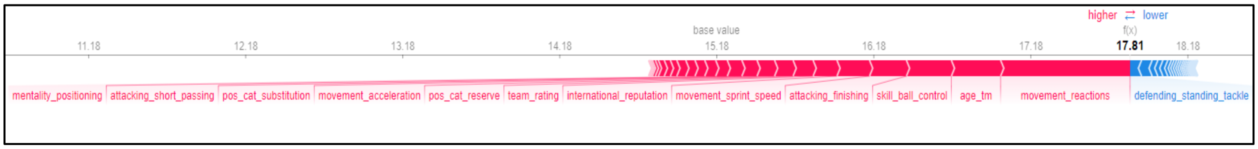


Figure 5-11: Timo Werner Prediction — SHAP values

In Figure 5-10, the model predicts 18.70 (a log value), whereas the base market value is 15.18 (the average of all predictions). Features increasing the model output (i.e., predicted market value) are in red color, and their visual size (i.e., arrow length) shows the magnitude of the feature's effect. Features decreasing the model output are in blue color. The biggest impact comes from age. Mbappé was only 21 years old in 2019/2020 season. The model also recognizes his athleticism (e.g., reactions, sprint speed, acceleration, stamina) and position as an attacking player (e.g., finishing, short pass). Team rating is an important feature. Both Mbappé and Neymar play PSG in French Ligue 1, a rising power that made it to the final of the Champion League in the past 2019/2020 season. The standing tackle value has an effect decreasing the prediction. For attacking players like Mbappé, defensive skills like standing tackle are not necessary skills. Mbappé is a key player of PSG, which means he has plenty of playing time. The distance from the base value to the output value (i.e., model prediction) equals total length of the blue arrows minus the total length of the red arrows.

In Figure 5-11, the model predicts 17.81 (a log value). In contrast to Mbappé's case, team rating has a moderate positive impact on Timer Werner's predicted market value. Timo Werner plays for RB Leipzig in German Bundesliga, which is an underdog team in the Champion League. Werner is not supported by particularly potent teammates. The model also acknowledges Werner's athleticism and attacking skills. His age has a smaller positive effect than Mbappé. Werner is just two years elder and is akin to Mbappé in terms of their repertoire of a forward, but his market value is much lower than Mbappé according to the model prediction. For clubs could not afford Mbappé's transfer fee, Werner might be a more realistic target that also makes more economic sense.

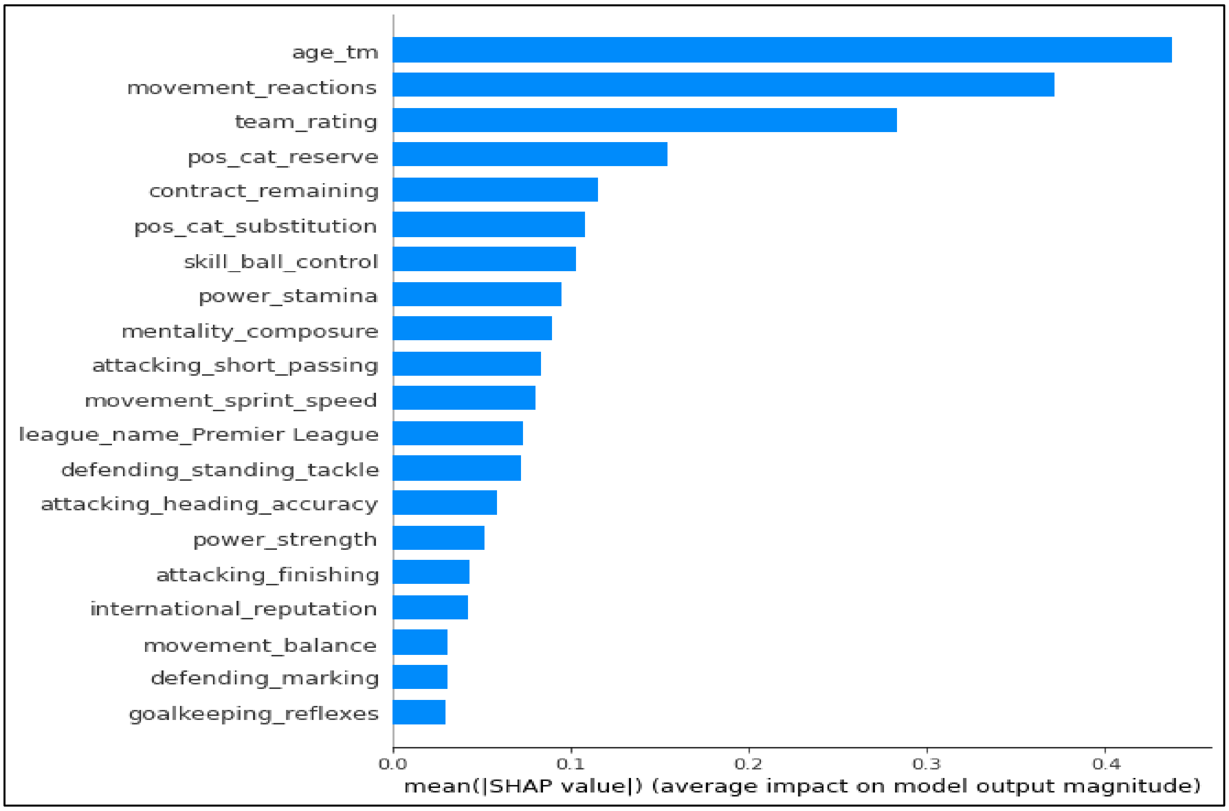


Figure 5-12: Standard SHAP Values Bar Plot

Features with large absolute Shapley values are important (Molnar, 2021). Figure 5-12 is a standard SHAP value bar chart that gives a notion of relative feature importance, but they do not provide the spread and distribution of impacts that feature has on the model's output, and how the feature's value relates to its impact (Lundberg et al., 2019). SHAP summary plots replace typical bar charts of global feature importance.

SHAP summary plots compute an entire dataset to explain a model's expected overall behavior (global), i.e., how the model behaves on average for a given dataset. Features are first sorted by their global impact, $\sum_{j=1}^{N} \left|\phi_i^{(j)}\right|$. Vertical axis shows what feature it is depicting. The features are ordered according to their importance. Each dot on the summary plot represents a SHAP value $\phi_i^{(j)}$for a feature i of an instance j. Dots representing the same feature are plotted horizontally, stacking vertically when there is not enough space. Those overlapping dots give a sense of the distribution (density) of the SHAP values per feature (Molnar, 2021, "5.10.6 SHAP Summary Plot", para. 1). Clusters generally mean similar feature values have comparable effects on the model output. Colors show whether the values of a feature are high or low. Blue color indicates low feature value whereas pink color shows high feature value. Horizontal location shows whether the effect of that value caused a higher or lower prediction. For example, on the left side, a negative SHAP value means a negative effect on the target variable.

In Figure 5-13, age is the most important feature. Old age, represented by red dots, evidently has a negative SHAP value, as nearly all these red dots are plotted on the left side of the graph. The negative SHAP values of old age translate into a negative impact on the predicted market value. The age impact on the predicted market value varies smoothly as this coloring also has a smooth gradation. For players whose age are at the similar level, the age effect could vary greatly. The general trend of long tails reaching to both left and right means that extreme values (outliers) of age can significantly raise or lower market value. Reaction is the second most important feature. Again, the coloring shows a smooth increase in the model's output (a log value) as the feature value of reaction increase. The similar pattern holds for the team rating feature. Playing for a competitive team (i.e., high team rating) translates into a positive effect on market value. The value of team rating feature is proportional to its effect on market value. Position category reserve is a dummy variable. For players with reserve status (*pos_cat_reserve* = 1), this has a markedly negative effect on market value, though dots are not clustered around the left. For players with non-reserve status (*pos_cat_reserve* = 0), dots form a cluster to show similar yet relatively low positive effect. Position category substitution is also a dummy variable but forms two symmetrical clusters. Being a substitution player has a much lower negative effect than a reserve player. Not being a substitution player cannot significantly raise the predicted market value. Substitution and reserve status in a squad and can be thought of as indicators of playing time, which could indirectly test the effect of playing time on market value. Playing for an EPL club translates into a positive effect on market value. Not playing for an EPL club only has a minor negative effect. playing for an EPL team and his England national status have very mild effects on the predicted market value than the other sports science factors. As for contract remaining, the more years remaining in the contract, the more positive effect on market value. International reputation only has a detectable impact for a minority of players with high reputation. In contrast, stamina and composure could contribute to a relatively small positive impact for a majority of players with high value of either feature, but a low feature value could substantially lower the predicted market value. Balance and marking are not very informative in a sense that there is no interesting patter.

A more detailed interpretation of an influence of a single feature on the model output is given by SHAP dependent contribution plots. Figure 5-14, Figure 5-15, Figure 5-16 represent the influence of age, contract remaining, and international reputation respectively. All three features are not on their original scale, since robust scaling has been applied. But the general trends remain unaffected. As age increases, its SHAP value steadily declines from a positive number to a negative number. For players with one year remaining in the contract, the negative effect on market value is enormous for some of them but is almost negligible (SHAP values near 0) for others. International reputation has a negative albeit rather minor effect for little-known and lesser-known players. By contrast, reputation effect is notably positive for well-known players. The effect of having a certain value of each of three features is not constant. Instead, it varies a lot depending on the value of other features.

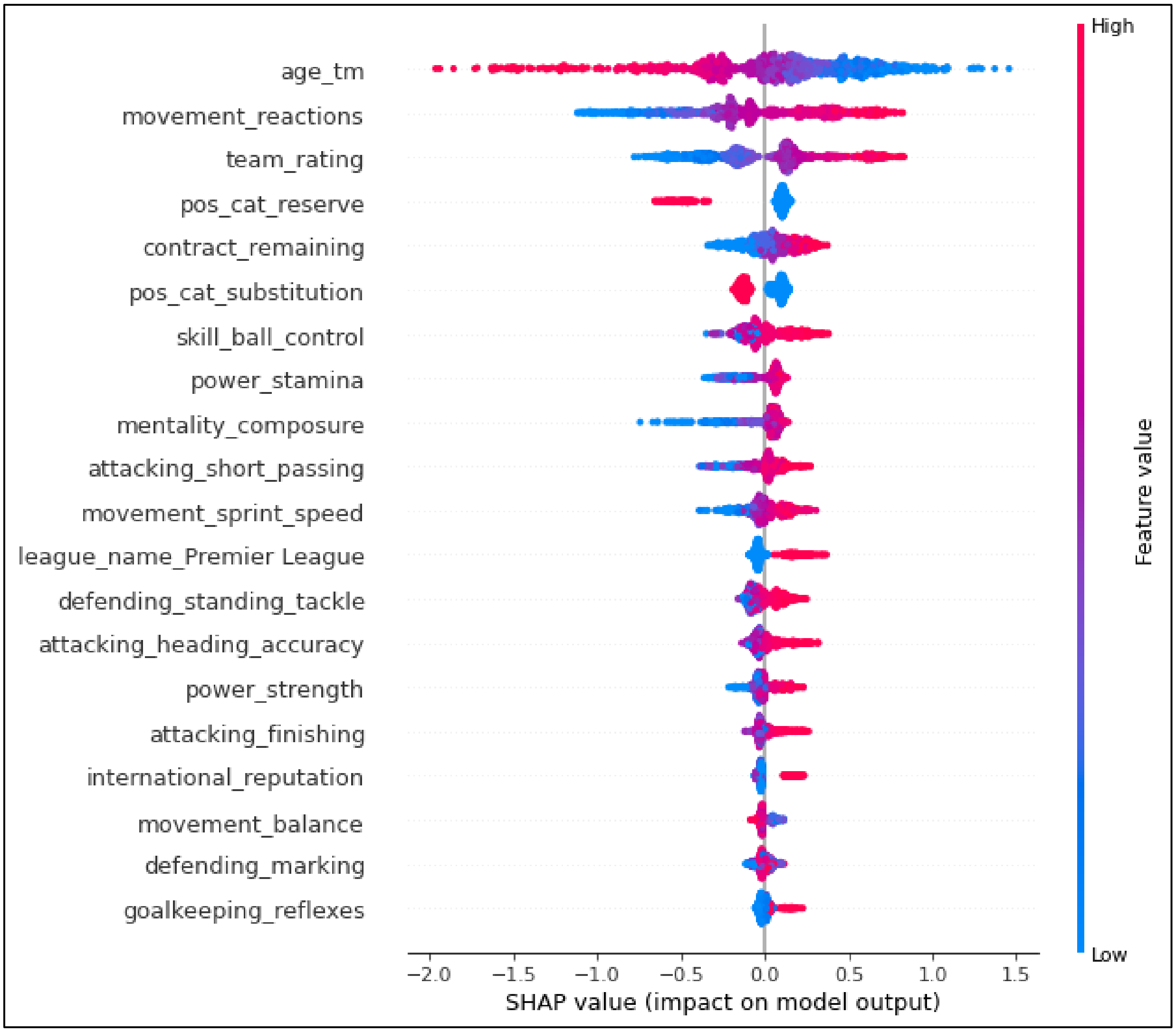


Figure 5-13: XGBoost Model Feature Attribution — SHAP Values Summary Plot

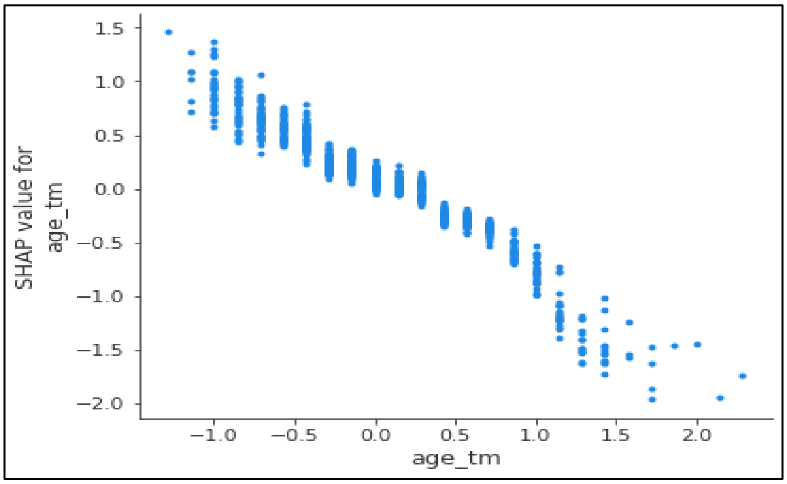


Figure 5-14: Age Dependence Plot

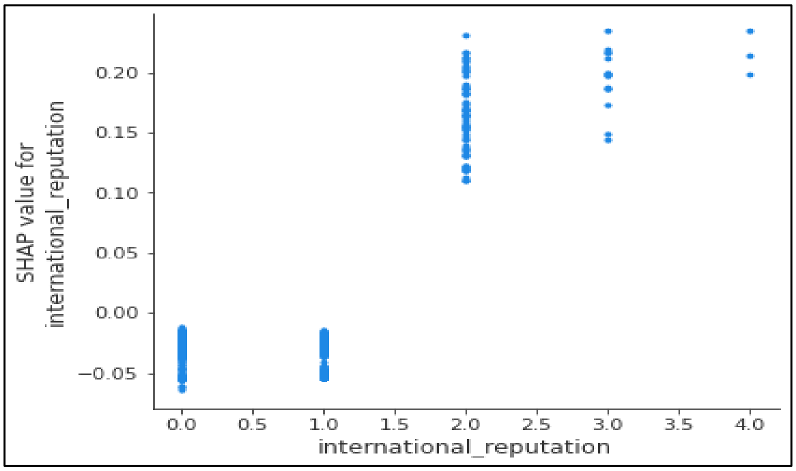


Figure 5-15: International Reputation Dependence Plot

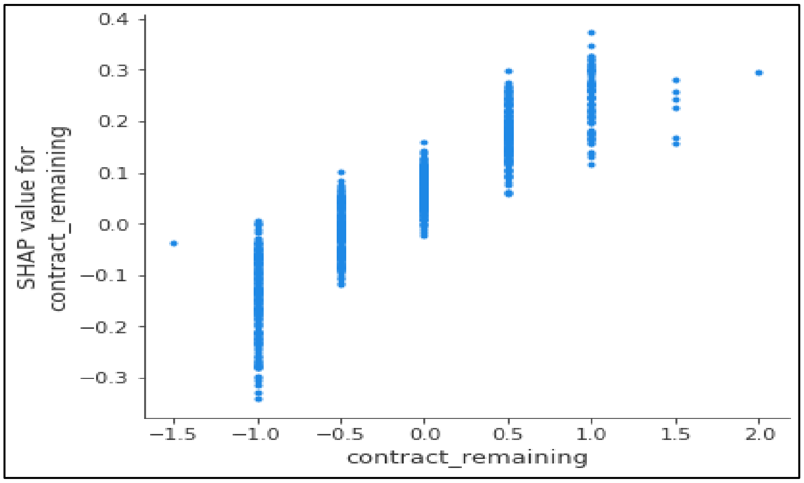


Figure 5-16: Contract Remaining Dependence Plot

Table 5-4 and Table 5-5 are a selection of market value predictions with low margin and high margin respectively, ranked by market value on Transfermarkt.com. In Table 5-4, for example, the predicted market value of Raheem Sterling and Niklas Süle is almost identical to their market value on Transfermart.com. In the 2020 summer, Timo Werner moved from RB Leipzig to Chelsea at a transfer fee believed to be in the region of €53M (Olley, 2020), which is very close to the model output (€54M). In Table 5-5, the predicted market value of Kylian Mbappé is much lower than his market value on Transfermart.com. Transfermarkt.com gave Mbappé a valuation of € 180,000,000 partly because he is widely believed to huge potential to reach the same level of Messi and Ronaldo. However, player potential, as a vague construct, may not be well represented by the existing features (e.g., age). That is to say, Mbappé's potential may be a missing piece of information in the data that gives rise to the high margin. High margin or low margin predictions can be deconstructed by SHAP values, so researchers and managers obtain fine-grained comparisons.

Table 5-4: Top Low Margin Predictions

Player	Club	Market Value	Predicted Market Value	Margin	Margin Percentage
Raheem Sterling	Manchester City	€ 128,000,000	€ 127,074,968	925032	1%
Sadio Mané	Liverpool	€ 120,000,000	€ 127,044,688	-7044688	-6%
Leroy Sané	Manchester City	€ 80,000,000	€ 71,764,712	8235288	10%
Joshua Kimmich	FC Bayern Munich	€ 64,000,000	€ 55,838,732	8161268	13%
Raphaël Varane	Real Madrid	€ 64,000,000	€ 52,281,236	11718764	18%
Timo Werner	RB Leipzig	€ 64,000,000	€ 54,540,864	9459136	15%
Miralem Pjanić	Juventus	€ 52,000,000	€ 44,954,460	7045540	14%
Mauro Icardi	Paris Saint-Germain	€ 52,000,000	€ 47,916,152	4083848	8%
Niklas Süle	FC Bayern Munich	€ 48,000,000	€ 48,292,516	-292516	-1%
Koke	Atlético Madrid	€ 48,000,000	€ 53,316,972	-5316972	-11%

Table 5-5: Top High Margin Predictions

Player	Club	Market Value	Predicted Market Value	Margin	Margin Percentage
Kylian Mbappé	Paris Saint-Germain	€ 180,000,000	€ 131,999,288	48000712	27%
Harry Kane	Tottenham Hotspur	€ 120,000,000	€ 88,287,936	31712064	26%
Lionel Messi	FC Barcelona	€ 112,000,000	€ 84,116,328	27883672	25%
Virgil van Dijk	Liverpool	€ 80,000,000	€ 53,531,468	26468532	33%
N'Golo Kanté	Chelsea	€ 80,000,000	€ 37,272,180	42727820	53%
Roberto Firmino	Liverpool	€ 72,000,000	€ 94,878,240	-22878240	-32%
Paulo Dybala	Juventus	€ 72,000,000	€ 100,405,968	-28405968	-39%
Alisson Becker	Liverpool	€ 72,000,000	€ 53,349,012	18650988	26%
Dele Alli	Tottenham Hotspur	€ 64,000,000	€ 87,883,704	-23883704	-37%
Andrew Robertson	Liverpool	€ 64,000,000	€ 38,048,508	25951492	41%

Chapter 6 Discussion and Implications

In competitive soccer leagues where all clubs are operating under financial constraints, player valuation is a high-stakes decision for the franchises. Sound investments in the transfer market immensely benefit club operations, while personnel mistakes consign a team to deplorable performance. Recent practices have shown clubs' growing interest in the use of data and analytics. Managers and sporting directors need new technologies to screen target players, estimate their market values, and shape acquisition practices. Executives want a robust mitigation strategy for implicit or even systematic human biases in the valuation process. Addressing inefficiencies in the transfer market has granted clubs a thin edge that might convert into more wins per million dollars (win efficiency) and additional franchise value through corporate sponsorship and television contracts (Valerdi, 2017). In broad human resources management, the adoption of emerging technologies such as computer vision is part of the latest trend towards more machine-based skills assessments and talent recruitment (West, 2021). With automation pervading many fields, a future of human soccer managers collaborating with AI-powered assistants may be within reach (West, 2021).

A few key managerial and economic implications emerge from this study. First, a small number of economic and risk factors have larger impacts on player valuation than numerous sports science factors. Table 6-1 is a selection of some key drivers of player valuation. Age and contract remaining are of great economic relevance. Put it another way, many sports science factors, at an individual level, might not be as important as a few fundamental economic and risk factors. The conventional wisdom may give too much weight on the soccer skill aspect in the valuation process. Age is a negative value driver of singular importance. Not only do soccer clubs value a player's current form, but also invest in his room for improvement in years to come. Contract remaining being a value driver is in line with the literature. Multiple years remaining in the contract give a player's incumbent club more bargaining power when negotiating buyer clubs. Players who recently signed or renewed a long-term contract would receive higher valuation compared with their cohorts who have one or two years left in the contract, as buyers cannot exploit a potential free transfer as the leverage to negotiate the transfer fee. Substitution status is a negative value driver. Substitution players generally have fewer playing opportunities than key players. The high frequency of being substituted and the lack of appearances have economic repercussions. Even a good player who has not been given enough minutes played would witness a drop in his market value (e.g., Gareth Bale). In this work, players who do not fall into substitution or reserve category typically have sufficient playing time, irrespective of their general position category (e.g., attacker, midfielder, defender). Belonging to EPL is a value driver, as it implies the importance of environment in the valuation process. EPL is widely regarded as the most competitive and richest league. Playing for a club in EPL is a boon from a valuation standpoint. International reputation is a value driver in a sense that it is a proxy for player popularity. Interestingly, data analysis of this work does not lend credence to the claims of certain nationalities being a key value driver, although German and French nationality have very mild positive effect on market value (See Figure A-2 in Appendix). Germany and France are the champions of the past two world cups. Neither England nor Brazilian national has a major effect on player valuation, despite that the price premium of players from both countries have been well documented in the past.

Second, within sports science factors, general physiological (e.g., reactions, stamina, sprint speed, strength) and psychological attributes (e.g., composure) appear to be more important than many soccer-specific skills (e.g., finishing). Unlike finishing or passing, reactions and stamina are physiological constructs that are not very soccer specific. Many individual and team sports demand quick reactions and enough stamina. Endurance has become an essential characteristic for soccer players to thrive in high-intensity drills and games. The distance a player has covered during a game sometimes exceeds 10 km (Smith, 2020), as the latest trend in soccer tactics prioritizes off-ball movement. In soccer, the most prominent statistic is goals scored, but finishing ability is only one value driver, even less important than many abovementioned drivers. It would be unwise to judge Andrés Iniesta on how many goals he scored in a season. That said, had soccer skills, at a collective level, not been included in player valuation, the ability of ML models to produce accurate market value estimations would have diminished.

Table 6-1: Value Driver Selection

SHAP Value Rank	Value Driver	Positive/Negative Value Driver	Group
1	Age	Negative	**Economic and Risk Factor**
5	Contract Remaining	Positive	
6	Substitution Status	Negative	
12	Premier League	Positive	
17	International Reputation	Positive	
2	Reaction	Positive	**Physiological and Psychological Attribute**
8	Stamina	Positive	
9	Composure	Positive	
11	Sprint Speed	Positive	
15	Strength	Positive	
7	Ball Control	Positive	**Soccer Skill**
10	Short Passing	Positive	
13	Standing Tackle	Positive	
14	Heading Accuracy	Positive	
16	Finishing	Positive	

Third, this work has implications for the adoption of analytics in the business settings where model accuracy and interpretability are equally important. Machine learning provides a new approach to study sports economy, finance, and management. As with many application domains, large-scale data and machine learning algorithms help invent efficacious decision support tools for player valuation. The common pitfall of many ML-based models is the lack of explanation behind prediction results. The use of SHAP values and other visualizations can redress this situation and expand the benefits of analytics in sports management and other similar undertakings. For example, SHAP values quantify the marginal contributions of the key drivers to a player’s market valuation, which enhances interpretability of clubs’ strategic decision-making. Furthermore, while data analytics has become essential for affluent soccer clubs, more clubs with small budgets begin to adopt it by virtue of cheaper and more accessible software (Harper, 2021). This work’s quantitative analyses entirely rely on publicly available data (Transfermarkt.com and Sofifa) and are performed by free, open-source analytics software.

Sports gaming (e-Sports) data is a relatively new source of acquiring measurements of physiological and psychological attributes and soccer skill that are otherwise hard to measure in a real-world or experiment setting. Sports gaming itself is a data-driven industry. Soccer video games like EA FIFA, Pro Evolution Soccer, and Football Manager, have strong research and development departments that routinely collect and analyze massive datasets to simulate real-world soccer operations with high fidelity. The utilization of those datasets can propel research in sport business fields. This work has demonstrated not only the power of analytics in processing complex open-source data, but also the usage of interpretable ML techniques to make sense of high-stakes business decisions.

Lastly, collaborations between video analysts, sports scientists, business researchers, and practitioners may hold the key to extract value from every bit of information in complex soccer datasets. From a soccer manager's point of view, a synergy of analytics capabilities and traditional human scouting is the path forward, especially when combined with informative data visualization, storytelling and model interpretation techniques. Most of soccer research using machine learning approaches were conducted by computer science or electrical engineering researchers. Business researchers should be able to not only harness the power of machine learning but also to have some basic understanding of its theoretical and computational underpinnings. The introduction of data technologies also requires a tight interchange with practitioners and a discussion of how to share data and techniques within the research community (Rein & Memmert, 2016). Given the advent of analytics in sports that occurred in the last decade, soccer provokes current epistemological debates regarding the use (and primacy) of quantitative versus qualitative data in management and decision science research.

Chapter 7 Conclusion, Limitation and Future Research

7.1 Conclusion

Ever since its manifesto in England in the middle of the 19th century, soccer has undergone monumental transformations into professionalization and commercialization. The current waves of data analytics have intensified its evolution, as disruptive technologies are increasingly embedded in a host of soccer activities beyond mere performance enhancement. Despite such trends, soccer analytics within business disciplines is understudied. The motivations of this thesis are threefold: 1) sports analytics has significant managerial implications for business value creation; 2) modeling continuous team sports such as soccer is challenging given their mathematical nature; 3) emerging technologies help sports institutions invent instruments and strategies that would otherwise be impossible (e.g., tracking data). With these motivations, this thesis aims at addressing an overarching research question of theoretical and empirical importance: *what are the key drivers of player valuation in the soccer transfer market?*

To this end, this paper begins with a brief introduction of the genesis of Moneyball, the state of the art of analytics in the sports industry, and a characterization of Moneyball as a philosophical approach that not only considers distinctive properties of soccer but also represents transferable knowledge. Then, this thesis systematically reviews economics, finance, and sports science literature as theoretical foundations for soccer player valuation. Building upon the literature review, the thesis proposes a conceptual framework of soccer player valuation that integrates key attributes from business research and sports science. To the best of my knowledge, this integrated framework is the first attempt to highlight the interdisciplinarity of player valuation as well as the applicability of economic and financial theories to player valuation. Economic theories address market efficiency and equilibrium conditions. In a fully efficient transfer market, player valuation should absorb all relevant information. Equilibrium conditions explain 1) how economic factors would drive the supply and demand curves; 2) how the actual transfer fees of players are negotiated between the buying club and the selling club. Financial theories provide appropriate pricing frameworks. The hedonic pricing theory establishes the service of a soccer player as a hedonic product of which the price (e.g., market value) is an aggregate function of all utility generating attributes. The option pricing theory incorporates risk and uncertainty into the valuation process. Sport science, as an additional lens, substantially expands the selection of utility generating attributes (i.e., physiological and psychological attributes and soccer skills). This proposed framework could also serve as a guideline for player valuation in other sports or player wage projection.

Next, this thesis operationalizes the proposed conceptual framework by using computational modeling methods to learn from consolidated historical soccer data. XGBoost, a state-of-the-art ML algorithm, learns a reasonably good approximation of the player valuation function with the lowest RMSE and highest adjusted R^2. SHAP values, an interpretable ML method, identify the value drivers by quantifying the effect of each feature on market value and explain the predictions made by the best XGBoost model both on a collective and on an individual level. Specifically, SHAP values extract and visually depicts both the features that most contributed to market value and those that offset it, which is a handy tool at managers' disposal to enhance the transparency of the valuation process. Contrary to the conventional wisdom, SHAP values reveal that 1) a few fundamental economic and risk factors appear to be more important than a large number of sports science factors; 2) general physiological and psychological attributes seem to be more important than many soccer-specific skills. The proposed technological stack has shown the benefits of integrating ML technologies into mainstream sports business research.

7.2 Limitations and Future Research

Soccer epitomizes a sport that is hard to measure the fairness of any price or to weed out the subjectivity in the human cognitive processes. This thesis offers some remarks on the limitations of the present work and possible future directions in extending the proposed approach. The first limitation has to do with the moderate sample size (data volume). All players of the sample are from the "big 5" league in two consecutive seasons (2018/2019 and 2019/2020). If the model learns from a massive number of players in the training phase, it will generate more precise predictions. To enrich the sample, researchers can collect market value data from second tier leagues (e.g., Dutch Eredivise, Portuguese Primeira Liga) and expand the training data to more seasons. Data integration challenges also refrain this work from including more advanced performance metrics from new sources (e.g., xG, and xA at whoscored.com[7]). For instance, it is cumbersome to join market value data and player attributes data from different sources without a shared unique identifier (e.g., player ID).

The second limitation is data veracity (quality). This work emphasizes two squad statues that are associated with less playing time, namely, substitution and reserve. However, it is difficult to develop a specialized model for each general player position category such as attackers, midfielders, and defenders, as the datasets do not differentiate player position from squad status. To enrich samples in each position category, research needs a more coherent player position classification schema. In fact, to identify a player's best or most common position can be a classification task as such. For example, Pappalardo et al (2019) used unsupervised learning (clustering) to automatically identify player positions. Likewise, player popularity is an ever-changing, multi-faceted construct that can be measured in different ways. International reputation is just a convenient measurement of popularity and by no means the best one. The magnitude of popularity on player valuation warrants further scrutiny.

The third limitation is data variety. For example, the team rating feature is a rudimentary measurement of team strength, as it takes the average of individual players' overall rating within a team. Soccer Power Index (SPI) made by forecasting website FiveThirtyEight is a potentially better alternative for team rating in that it takes match results into account. This work does not include the historical data of clubs' actual transfer income and expenditure. Player chemistry is another dimension that should have been part of the valuation process. In a soccer team, cooperation, coordination, and complementation between players are the key to succeed (Al-Madi et al., 2016). Instructive in this respect is the study of player chemistry by Bransen (2020). A wealth of contextual or augmented event data has been harvested through cutting-edge tracking technologies, which enables more refined analysis. However, the sheer amount of data might become an obstacle in itself (Rein & Memmert, 2016). For example, as different kinds of data (e.g., market value data and tracking data) converge, researchers may be confronted with trade-offs between data variety and veracity.

Even with the literature flourishing, scholars cannot fully explain the irregularities of the transfer market (Kroken & Hashi, 2017). This thesis has established a discourse on sports

[7] https://www.whoscored.com/

analytics within the business disciplines, inspiring future studies in player valuation. First, future research could propose a composite popularity score that combines players google search trend and social media followers. The convergence of technology and sports has gained momentum as professional leagues strive to connect with millennial fans via social media (Valerdi, 2017). Players, as online celebrities, amass a huge following on social media platforms (e.g., Facebook, Twitter, Instagram, TikTok). The social media artifacts they created could be used to keep track of their publicity and media exposure. Second, future research could better assess player characters (mentalities). Sporting director Luis Campos has partnered the top university psychology department to profile players and has unearthed over £500m worth of soccer talents such as Kylian Mbappé during his tenure at Monaco and Lille (Sky Sports, 2019). Social media analytics can help clubs observe the mentality of potential targets. For example, sentiment analysis and topic modeling can be applied to target players' social media posts. If a player often sends aggressive tweets, there is a chance that he is not mentally stable. If his Instagram feeds show he is often at parties or on a luxury yacht or in a night club, his self-discipline and professionalism might be a concern. Traditional human scouting still offers valuable input, particularly for players who are unknown to the coaches' and scouts' networks (Nalton, 2020). Soccer consultants sometimes even commission dossiers on players' personalities. With appropriate data governance and privacy policies, these non-structured data sources could further improve predictive models. Third, the predictive analytics approach of this work could be extended to a next generation, utility-based player recommender system framework which builds a diversified portfolio of talents for clubs to tap into. In intelligent agents, utility-based agents can measure how desirable a particular state is (Zihayat et al., 2019). In a utility-based player recommendation framework, a utility function can be defined such that how competitive a club (i.e., agent) will be if it moves to a particular state (i.e., acquisition of a specific player). With all that said, soccer very much remains a game of human ingenuity. Machine learning does not supersede qualitative methods, common sense or expert knowledge but rather collaborates with them in a symbiotic way.

References

A Guide to Sabermetric Research. Society for American Baseball Research. (n.d.). Retrieved June 11, 2020, from https://sabr.org/sabermetrics.

Aarshay, J. (2016, March 1). Complete Guide to Parameter Tuning in XGBoost with codes in Python. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

Adler, M. (1985). Stardom and Talent. *The American Economic Review*, 75(1), 208–212. Retrieved April 14, 2021, from http://www.jstor.org/stable/1812714

Alamar, B., & Mehrotra, V. (2011a). Beyond 'Moneyball': Rapidly evolving world of sports analytics, Part I. *Analytics Magazine*. http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i/

Alamar, B., & Mehrotra, V. (2012). Analytics & Sports, Part III: Improving resource allocation with portfolio decision analysis. *Analytics Magazine*. http://analytics-magazine.org/analytics-a-sports-part-iii-improving-resource-allocation-with-portfolio-decision-analysis/

Ali, A. (2011). Measuring soccer skill performance: a review. *Scandinavian Journal of Medicine & Science in Sports*, *21*(2), 170-183. https://doi.org/10.1111/j.1600-0838.2010.01256.x

Allen, M. (2018, February 8). Analyzing the efficiency of transfer markets. *Chance Analytics*. https://chanceanalytics.wordpress.com/2018/02/08/analysing-the-efficiency-of-transfer-markets/

Al-Madi, F., Al-Tarawneh, K. I., & Alshammari, M. A. (2016). HR Practices in the Soccer Industry: Promising Research Arena. *International Review of Management and Marketing*, *6*(4), 641-653.

Amir, E., & Livne, G. (2005). Accounting, Valuation and Duration of Football Player Contracts. *Journal of Business Finance & Accounting*, *32*(3–4), 549–586. https://doi.org/10.1111/j.0306-686X.2005.00604.x

Anonymous. (2017, August 7). How a football transfer works? *The Economist*. https://www.economist.com/the-economist-explains/2017/08/07/how-a-football-transfer-works

Anonymous. (2019, October 8). Luis Campos: The Transfer Chief with the Midas Touch. *Sky Sports*. https://www.skysports.com/football/news/11095/11830121/luis-campos-the-transfer-chief-with-the-midas-touch

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2020). Explaining Anomalies Detected by Autoencoders Using SHAP. ArXiv:1903.02407 [Cs, Stat]. http://arxiv.org/abs/1903.02407

Baca, A. (Ed.). (2014). *Computer Science in Sport: Research and Practice*. (1st ed.). Routledge.

Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., & Van Gael, J. (2012). Crowd IQ: aggregating opinions to boost performance. *Proceedings of the 11th international conference on autonomous agents and multiagent systems*, *1*, 535–542.

Barnsley, R. H., Thompson, A. H., & Legault, P. (1992). Family Planning: Football Style. The Relative Age Effect in Football. *International Review for the Sociology of Sport*, *27*(1), 77–87. https://doi.org/10.1177/101269029202700105

Belson, K. (2020, April 16). Can't Scout Players in Person? The N.F.L. Turns to a Brooklyn Start-Up. *New York Times*. https://www.nytimes.com/2020/04/16/sports/football/nfl-draft-scouting-technology.html

Berg, E. W. A. van den. (2011). The Valuation of Human Capital in the Football Player Transfer Market: An investigation of transfer fees paid and received in the English Premier League. Chapter I – The Context of the Football Transfer Market. 65.

Blockeel, H., & Vanschoren, J. (2007). Experiment Databases: Towards an Improved Experimental Methodology in Machine Learning. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Knowledge Discovery in Databases: PKDD 2007* (6–17). Springer Berlin Heidelberg.

Bradley, P. S., Carling, C., Gomez Diaz, A., Hood, P., Barnes, C., Ade, J., Boddy, M., Krustrup, P., & Mohr, M. (2013). Match performance and physical capacity of players in the top three competitive standards of English professional soccer. *Human Movement Science*, *32*(4), 808–821. https://doi.org/10.1016/j.humov.2013.06.002

Bransen, L. (2020). Player Chemistry: Striving for a Perfectly Balanced Soccer Team. *2020 MIT Sloan Sports Analytics Conference*. http://www.sloansportsconference.com/wp-content/uploads/2020/02/Bransen_paper_player_chemistry.pdf

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Brownlee, J. (2016, August 31). Feature Importance and Feature Selection with XGBoost in Python. *Machine Learning Mastery*. https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/

Bryson, A., Frick, B., & Simmons, R. (2013). The Returns to Scarce Talent: Footedness and Player Remuneration in European Soccer. *Journal of Sports Economics*, *14*(6), 606–628. https://doi.org/10.1177/1527002511435118

Box, G. E. P., & Draper, N. R. (1987). *Wiley series in probability and mathematical statistics*. Empirical model-building and response surfaces. John Wiley & Sons.

Buekers, M., Borry, P., & Rowe, P. (2015). Talent in sports. Some reflections about the search for future champions. *Movement & Sport Sciences - Science & Motricité*, 88, 3–12. https://doi.org/10.1051/sm/2014002

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

Burton, D., & Naylor, S. (1997). Is anxiety really facilitative? reaction to the myth that cognitive anxiety always impairs sport performance. *Journal of Applied Sport Psychology*, *9*(2), 295-302. https://doi.org/10.1080/10413209708406488

Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: Theory and evidence. *Applied Economics*, *25*(12), 1467–1476. https://doi.org/10.1080/00036849300000150

Carmichael, F., Forrest, D. & Simmons, R. (1999). The labour market in association football: Who gets transferred and for how much? *Bulletin of Economic Research*, *51*(2), 125-150.

Caya, O., & Bourdon, A. (2016). A Framework of Value Creation from Business Intelligence and Analytics in Competitive Sports. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Koloa, HI, USA, 1061–1071. https://doi.org/10.1109/HICSS.2016.136

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cokins, G., DeGrange, W., Chambal, S & Walker, R. (2016). Sports Analytics Taxonomy, V1.0. *ORMS Today*. https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/Sports-analytics-taxonomy-V1.0

Coluccia, D., Fontana, S., & Solimene, S. (2018). An application of the option-pricing model to the valuation of a football player in the "Serie A League." *International Journal of Sport Management and Marketing*, *18*(1/2), 155. https://doi.org/10.1504/IJSMM.2018.091345

Conn, D. (2017). Premier League remains world's richest courtesy of huge TV revenue growth. *The Guardian*. https://www.theguardian.com/football/2017/jul/11/premier-league-worlds-richest-tv-revenue-growth

Daumé, H. (2017). A Course in Machine Learning. http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf

Davenport, T. H. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business School Press.

Davenport, T.H. (2011). Six Things Your Company Has in Common with the Oakland A's. *Harvard Business Review*. https://hbr.org/2011/09/six-things-your-company-has-in

Davenport, T.H. (2014a). What businesses can learn from sports analytics. *MIT Sloan Management Review*, *55*(4), 10-13.

Davenport, T.H. (2014b). Analytics in Sports: The New Science of Winning. https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf

Denny, K., & Sullivan, V. O. (2007). The Economic Consequences of Being Left-Handed Some Sinister Results. *Journal of Human Resources*, *42*(2), 353–374. https://doi.org/10.3368/jhr.XLII.2.353

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

DW Kick off! (2019, October 15). *WHY lefties are better footballers*. YouTube. https://youtu.be/roEsJo7kdOQ

Fama, E. F. (1970). Efficient Capital Markets - Review of Theory and Empirical Work. *Journal of Finance*, *25*(2), 383-423. http://www.jstor.com/stable/2325486

Fernández, J., Bornn, L., & Cervone, D. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *2019 MIT Sloan Sports Analytics Conference*. http://www.sloansportsconference.com/wp-content/uploads/2019/02/Decomposing-the-Immeasurable-Sport.pdf

FIFA's transfer report: Big 5 accounted for almost three quarters of global spending in January 2020. FIFA. (2020b, February 12). https://www.fifa.com/who-we-are/news/fifa-s-transfer-report-big-5-accounted-for-almost-three-quarters-of-global-spend

Financial Fair Play. UEFA. (2019, June 5). https://www.uefa.com/insideuefa/protecting-the-game/financial-fair-play/

Flegl, M., Jiménez-Bandala, C. A., Lozano, C., & Andrade, L. (2018). Personnel selection in complex organizations: A case of Mexican football team for the 2018 World Cup in Russia. *Revista Del Centro de Investigación de La Universidad La Salle*, *13*(49), 43–66. https://doi.org/10.26457/recein.v13i49.1510

Florke, C. R., & Ecker, M. D. (2003). NBA Draft Lottery Probabilities. *American Journal of Undergraduate Research*, *2*(3), 19-29.

Franck, E., & Nüesch, S. (2012). Talent and/or Popularity: What Does It Take to Be a Superstar? *Economic Inquiry*, *50*(1), 202–216. https://doi.org/10.1111/j.1465-7295.2010.00360.x

Franks, A.M., D'Amour, A., Cervone, D., & Bornn, L. (2016). Meta-analytics: tools for understanding the statistical properties of sports metrics. *Journal of Quantitative Analysis in Sports*, *12*(4), 151-165. https://doi.org/10.1515/jqas-2016-0098

Frick, B. (2007). The Football Players' Labor Market: Empirical Evidence from the Major European Leagues. *Scottish Journal of Political Economy*, *54*(3), 422–446. https://doi.org/10.1111/j.1467-9485.2007.00423.x

Frick, B. (2011). Performance, Salaries, and Contract Length: Empirical Evidence from German Soccer. *International Journal of Sports Finance*, *6*, 87-118.

Fry T., Galanos, G., & Posso, A. (2014). Let's get Messi? Top-scorer productivity in the European Champions League. *Scottish Journal of Political Economy*, *61*(3), 261-279.

Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets - sport and economic contribution of Spanish soccer players. *Managerial and Decision Economics*, *28*(1), 57-70.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. (1st ed.). O'Reilly Media, Inc.

Gerrard, B. (2001). A new approach to measuring player and team quality in professional team sports. *European Sport Management Quarterly*, *1*(3), 219-234. https://doi.org/10.1080/16184740108721898

Gerrard, B. (2007). Is the Moneyball Approach Transferable to Complex Invasion Team Sports? *International Journal of Sport Finance*, *2*(4). https://search.proquest.com/docview/229399553?pq-origsite=gscholar

Gerrard, B. (2014). Achieving transactional efficiency in professional team sports: The theory and practice of player valuation. In J. Goddard & P. Sloane (Eds.), *Handbook on the Economics of Professional Football* (pp. 189-202). https://doi.org/10.4337/9781781003176

Gerrard, B. (2017). Analytics, Technology and High Performance Sport. In N. Schulenkorf & S. Frawley (Eds.), *Critical Issues in Global Sport Management*. Routledge.

Gianecchini, M., & Alvisi, A. (2015). Late Career of Superstar Soccer Players: Win, Play, or Gain? *30th EGOS Colloquium*.

Giuliani, M. (2012). The creation and destruction of value: the intellectual capital cycles. *Proceedings of the 4th European Conference on Intellectual Capital*, 212-219. http://academic-conferences.org/ecic/ecic2012/ecic12-home.htm.

Gladwell, M. (2011). *Outliers: The Story of Success*. Penguin Books.

González, Enric. "El Balón y la Bandera." El País (Madrid), May 31, 2008.

Harper, J. (2021, March 4). Data Experts Are Becoming Football's Best Signings. *BBC*. https://www.bbc.com/news/business-56164159

Harrison, O. (2018, September 10). Machine Learning Basics with the K-Nearest Neighbors Algorithms. *Medium*. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Herm, S., Callsen-Bracker, H.-M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review*, *17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006

Hoare, D. G. & Warr, C. R. (2000). Talent identification and women's soccer: An Australian experience. *Journal of Sports Sciences*, *18*(9), 751-758. https://doi.org/10.1080/02640410050120122

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, *108*(1), 29–47. https://doi.org/10.1007/s10994-018-5704-6

Huijgen B.C., Elferink-Gemser M.T., Lemmink K.A., Visscher C. (2014). Multidimensional performance characteristics in selected and deselected talented soccer players. *European Journal of Sport Science*, *14*(1), 2-10.

James, G., Witten, D., Hastie, T., & Tibshirani, Robert. (2017). *An Introduction to Statistical Learning: with Applications in R* (8th ed.). Springer.

Jones, P. D. (2015). Situation Awareness in Soccer. [Unpublished master's thesis]. Swansea University. http://cronfa.swan.ac.uk/Record/cronfa42481

Kahneman, D. (2012, June 15). Of 2 Minds: How Fast and Slow Thinking Shape Perception and Choice [Excerpt]. *Scientific American*. https://www.scientificamerican.com/article/kahneman-excerpt-thinking-fast-and-slow/

Kharrat, T., McHale, I. G., & Peña, J. L. (2019). Plus–minus player ratings for soccer. *European Journal of Operational Research*, *283*(2), 726-736. https://doi.org/10.1016/j.ejor.2019.11.026

Kim, Y., Bui, K.-H. N., & Jung, J. J. (2021). Data-driven exploratory approach on player valuation in football transfer market. *Concurrency and Computation: Practice and Experience*, *33*(3), e5353. https://doi.org/10.1002/cpe.5353

Kobielus, J. (2014, April 17). Moneyball is the true game-changing application of data analytics. (2014). IBM Big Data & Analytics Hub. https://www.ibmbigdatahub.com/blog/moneyball-true-game-changing-application-data-analytics

KPMG The European Elite 2020 Football Club's Valuation. (2020, May). footballbenchmark.com. https://footballbenchmark.com/documents/files/KPMG The European Elite 2020_Online version_.pdf

Kroken, C., & Hashi, G. (2017). Market efficiency in the European football transfer market. https://biopen.bi.no/bi-xmlui/handle/11250/2485695

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. (1st ed.). Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuper, S. & Szymanski, S. 2012. *Soccernomics* (2018 World Cup ed.). Harper Collins.

Lehmann, E. E., & Schulze, G. G. (2008). What Does it Take to be a Star? - The Role of Performance and the Media for German Soccer Players. *Applied Economics Quarterly*, *54*(1), 59–70. http://dx.doi.org/10.3790/aeq.54.1.59

LinkedIn Learning. (2013, October 17). *How did the Mavericks use analytics to beat the Miami Heat in the 2011 NBA Finals? | lynda.com*. YouTube. https://youtu.be/u3jY3TcCqGU

Louzada, F., Maiorano, A. C., & Ara, A. (2016). iSports: A web-oriented expert system for talent identification in soccer. *Expert Systems with Applications*, *44*, 400–412. https://doi.org/10.1016/j.eswa.2015.09.007

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. ArXiv:1802.03888 [Cs, Stat]. http://arxiv.org/abs/1802.03888

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. ArXiv:1705.07874 [Cs, Stat]. http://arxiv.org/abs/1705.07874

Magee W. (2017). Fans are already disconnected from their clubs—a summer of frenzied transfer hyperinflation only makes it worse. *The Independent*. https://www.independent.co.uk/sport/football/transfers/premier-league-summer-transfer-window-record-spending-net-spend-romelu-lukaku-paul-pogba-a7853226.html.

Majewski, S., & Majewska, A. (2017). Using Monte Carlo Methods for the Valuation of Intangible Assets in Sports Economics. *Folia Oeconomica Stetinensia*; Szczecin, *17*(2), 71–82. http://dx.doi.org/10.1515/foli-2017-0019

Martin, L. (2016). *Sports Performance Measurement and Analytics: The Science of Assessing Performance, Predicting Future Outcomes, Interpreting Statistical Models, and Evaluating the Market Values of Athletes*. Pearson Education, Inc.

Martin, L., & T. W. Miller. (2016). A Model for Measurement in Sports. Manhattan Beach, Calif.: Research Publishers. http://www.research-publishers.com/. 52

Massey, C., & Thaler, R. H. (2013). The Loser's Curse: Decision Making and Market Efficiency in the National Football League Draft. *Management Science*, *59*(7), 1479–1495. https://doi.org/10.1287/mnsc.1120.1657

Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S. L., & Ortega, G. J. (2018). Transfer market activities and sportive performance in European first football leagues: A dynamic network approach. *PLOS ONE*, *13*(12), e0209362. https://doi.org/10.1371/journal.pone.0209362

McDowall, M. (Producer), & McDowall, M. (Director). (2011). *Ronaldo: Tested to the Limit* [Castrol]. https://www.dailymotion.com/video/x3q2vew

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, Inc.

Molnar, C. (2021). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning—A Brief History, State-of-the-Art and Challenges. ArXiv:2010.09337 [Cs, Stat]. http://arxiv.org/abs/2010.09337

Mourao, P. R. (2016). Soccer transfers, team efficiency and the sports cycle in the most valued European soccer leagues – have European soccer teams been efficient in trading players? *Applied Economics*, *48*(56), 5513–5524. https://doi.org/10.1080/00036846.2016.1178851

Müller, A., Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, *263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005

Murphy, P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nalton, J. (2020, February 19). Soccer Analysis Moves Toward Smarter Scouting And More Accessible Data. *Forbes*. https://www.forbes.com/sites/jamesnalton/2020/02/19/soccer-analytics-smarter-scouting-and-more-accessible-data/?sh=6ec009131557

Norikazu Hirose. (2009). Relationships among birth-month distribution, skeletal age and anthropometric characteristics in adolescent elite soccer players. *Journal of Sports Sciences*, *27*(11), 1159-1166. https://doi.org/10.1080/02640410903225145

Nsolo E., Lambrix P., Carlsson N. (2019) Player Valuation in European Football. In: Brefeld U., Davis J., Van Haaren J., Zimmermann A. (Eds.), *Machine Learning and Data Mining for Sports Analytics*. MLSA 2018. Lecture Notes in Computer Science, *11330*. Springer, Cham. https://doi.org/10.1007/978-3-030-17274-9_4

Olley, J. (2020, June 18). Chelsea Agree Deal to Sign Timo Werner from RB Leipzig. *ESPN*. https://www.espn.com/soccer/soccer-transfers/story/4114910/chelsea-agree-deal-to-sign-timo-werner-from-rb-leipzig

Patnaik, D., Praharaj, H., Prakash, K., & Samdani, K. (2019). A study of Prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–7. https://doi.org/10.1109/ICSCAN.2019.8878843

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. *ACM Transactions on Intelligent Systems and Technology*, *10*(5), 59:1–59:27. https://doi.org/10.1145/3343172

Payyappalli, V. M., & Zhuang, J. (2019). A data-driven integer programming model for soccer clubs' decision making on player transfers. *Environment Systems and Decisions*, *39*(4), 466–481. https://doi.org/10.1007/s10669-019-09721-7

Pedace, R. (2008). Earnings, performance, and nationality discrimination in a highly competitive labor market as an analysis of the English professional soccer league. *Journal of Sports Economics*, *9*(2), 115-140.

Perciballi, S.G. (2011). Soccer and society: A study of ethnic group adaptation in society through the game of soccer; Windsor, Ontario, 1972. *Electronic Theses and Dissertations*, 272. https://scholar.uwindsor.ca/etd/272

Ployhart, R.E., Nyberg, A.J., Reilly, G., Maltarich, M.A. (2014), Human capital is dead; Long live human capital resources! *Journal of Management*, *40*(2), 371-398.

Poli, D. R., Ravenel, L., & Besson, R. (2020, March). CIES Football Observatory Monthly Report n°53—March 2020. 5. https://football-observatory.com/IMG/pdf/mr53en.pdf

Rao, A. R., & Bergen, M. E. (1992). Price Premium Variations as a Consequence of Buyers' Lack of Information. *Journal of Consumer Research*, *19*(3), 412–423. JSTOR.

Rastogi, S. K., & Deodhar, S. Y. (2009). Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty20 Vision. *Vikalpa*, *34*(2), 15–24. https://doi.org/10.1177/0256090920090202

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(Proc2). https://doi.org/10.14198/jhse.2017.12.Proc2.05

Regulations on the Status and Transfer of Players. FIFA. (2020a, June). https://resources.fifa.com/image/upload/regulations-on-the-status-and-transfer-of-players-june-2020.pdf?cloudid=ixztobdwje3tn2bztqcp

Reilly, T., Williams, A. M., Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences*, *18*(9), 695–702. https://doi.org/10.1080/02640410050120078

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1410. https://doi.org/10.1186/s40064-016-3108-2

Roach, M. A. (2018). Testing Labor Market Efficiency Across Position Groups in the NFL. *Journal of Sports Economics*, *19*(8), 1093–1121. https://doi.org/10.1177/1527002517704021

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34-55.

Rosen, S. (1981). The Economics of Superstars. *The American Economic Review*, *71*(5), 845–858. JSTOR.

Röthig, P., Prohl, R., & others. (2003. Sportwissenschaftliches Lexikon, Hofmann (available at http://www.ulb.tu-darmstadt.de/tocs/9970681.pdf).

Rottenberg, S. (2000). Resource Allocation and Income Distribution in Professional Team Sports. *Journal of Sports Economics*, *1*(1), 11–20

Rudin, C. (2018). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. ArXiv:1811.10154 [Cs, Stat]. http://arxiv.org/abs/1811.10154

Sandri, M. & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611–628.

Schoenfeld, B. (2019, May 22). How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory. *New York Times*. https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Serna Rodríguez, M., Ramírez Hassan, A., & Coad, A. (2019). Uncovering Value Drivers of High Performance Soccer Players. *Journal of Sports Economics*, *20*(6), 819–849. https://doi.org/10.1177/1527002518808344

Sethneha. (2020, November 9). Entropy – A Key Concept for All Data Science Beginners. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/11/entropy-a-key-concept-for-all-data-science-beginners/

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Shapiro, C (1983). Premium for High Quality Products as Returns to Reputations. *Quarterly Journal of Economics*, *98*(4), 659-79.

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, *35*(3), 553–572. JSTOR. https://doi.org/10.2307/23042796

Sæbø, O. D., & Hvattum, L. M. (2019). Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics*, *5*(1), 23–34. https://doi.org/10.3233/JSA-170235

Sierksma, G. (2006). Computer Support for Coaching and Scouting in Football. In E. F. Moritz & S. Haake (Eds.), *The Engineering of Sport 6* (215–219). Springer. https://doi.org/10.1007/978-0-387-45951-6_39

Sloane, P. J. (1971). Scottish Journal of Political Economy:the Economics of Professional Football: The Football Club as a Utility Maximiser*. *Scottish Journal of Political Economy*, *18*(2), 121–146. https://doi.org/10.1111/j.1467-9485.1971.tb00979.x

Sloane, P. J. (2015). The Economics of Professional Football Revisited. *Scottish Journal of Political Economy*, *62*(1), 1-7. https://doi.org/10.1111/sjpe.12063

Smith, A. (2020, May 28). Premier League Running Stats This Season Revealed: Will Premier League Players Match Pre-Break Fitness Levels? *Sky Sports*. https://www.skysports.com/football/news/11661/11996016/premier-league-running-stats-this-season-revealed

Soni, D. (2018, March 12). Introduction to k-Nearest-Neighbors. *Medium*. https://towardsdatascience.com/introduction-to-k-nearest-neighbors-3b534bb11d26

Stambulova, N., Stephan, Y., & Jäphag, U. (2007). Athletic retirement: A cross-national comparison of elite French and Swedish athletes. *Psychology of Sport and Exercise*, *8*(1), 101-118.

Stanojevic, R., & Gyarmati, L. (2016). Towards Data-Driven Football Player Assessment. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 167–172. https://doi.org/10.1109/ICDMW.2016.0031

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621. https://doi.org/10.1016/j.ijforecast.2010.01.003

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor Books.

Swanepoel, M. J., & Swanepoel, J. (2016). The Correlation between Player Valuation and THE Bargaining Position of Clubs in THE English Premier League (EPL). *International Journal of Economics and Finance Studies*, *8*(1), 17. https://repository.nwu.ac.za/handle/10394/24705

Szymanski, S. (2004). Professional Team Sports Are Only a Game: The Walrasian Fixed-Supply Conjecture Model, Contest-Nash Equilibrium, and the Invariance Principle. *Journal of Sports Economics*, *5*(2), 111–126. https://doi.org/10.1177/1527002503261485

Taussig, F.W. (2007). *Principles of Economics, 2*. Cosimo Inc.

Taylor, M.S., Giannantonio, C.M. (1993), Forming, adapting, and terminating the employment relationship: A review of the literature from individual, organizational, and interactionist perspectives. *Journal of Management*, *19*(2), 461-515.

Tran, U. S., & Voracek, M. (2016). Footedness Is Associated with Self-reported Sporting Performance and Motor Abilities in the General Population. *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01199

Tucker, R., & Collins, M. (2012). What makes champions? A review of the relative contribution of genes and training to sporting success. *British Journal of Sports Medicine*, *46*(8), 555–561. https://doi.org/10.1136/bjsports-2011-090548

Tunaru, R., Clark, E., & Viney, H. (2005). An option pricing framework for valuation of football players. *Review of Financial Economics*, *14*(3–4), 281–295. https://doi.org/10.1016/j.rfe.2004.11.002

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Valerdi, R. (2017). Why Software Is Like Baseball. *IEEE Software*, *34*(5), 7–9. https://doi.org/10.1109/MS.2017.3571583

Vallerand, R. J. 2004. Intrinsic and extrinsic motivation in sport. *Encyclopedia of Applied Psychology*, *2*(10), 52.

Vergeer, M., & Mulder, L. (2019). Football Players' Popularity on Twitter Explained: Performance on the Pitch or Performance on Twitter? *International Journal of Sport Communication*, *12*(3), 376–396. https://doi.org/10.1123/ijsc.2018-0171

West, D.M. (2021, March 18). How the NFL is using AI to evaluate players. *Brookings*. https://www.brookings.edu/blog/techtank/2021/03/18/how-the-nfl-is-using-ai-to-evaluate-players/?fbclid=IwAR2Xz5g6-ZDEcovkiZyRtDlQX7nFUL5IJCOVLmXY2opNiFc9pYaUdXwneKg

Williams, A.M. (2000). Perceptual skill in soccer: implications for talent identification and development. *Journal of Sports Science*, *18*(9), 737–750. https://doi.org/10.1080/02640410050120113

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Wright, P. M., Smart, D. L., & McMahan, G. C. (1995). Matches between human resources and strategy among NCAA basketball teams. *Academy of Management Journal*, *38*(4), 1052–1074. https://doi.org/10.2307/256620

Xiao, X., Chian Tan, F. T., Lim, E. T. K., Henningsson, S., Vatrapu, R., Hedman, J., Tan, C. W., Clemenson, T., Mukkamala, R. R., & Van Hillegersberg, J. (2018). Sports Digitalization: An Overview and A Research Agenda. *2017 38th International Conference on Information Systems (ICIS)*, Seoul, Republic of Korea. http://aisel.aisnet.org/icis2017/General/Presentations/6/

Yam, D. (2019). A Data Driven Goalkeeper Evaluation Framework. *2020 MIT Sloan Sports Analytics Conference*. http://www.sloansportsconference.com/wp-content/uploads/2019/02/Data-Driven-Goalkeeper-Evaluation-Framework-1.pdf

Yiu, T. (2019, June 12). Understanding Random Forest How the Algorithm Works and Why it Is so Effective. *Medium*. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Yorke, J. (2019, April 26). Pass Footedness in the Premier League. *StatsBomb*. https://statsbomb.com/2019/04/pass-footedness-in-the-premier-league/#:~:text=Right%20footed%20players%20average%20around,%2C%20left%20footed%20passes%2078.5%25.

Zheng, A. & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., & An, A. (2019). A Utility-Based News Recommendation System. *Decision Support Systems*, *117*, 14–27. https://doi.org/10.1016/j.dss.2018.12.001

Appendix

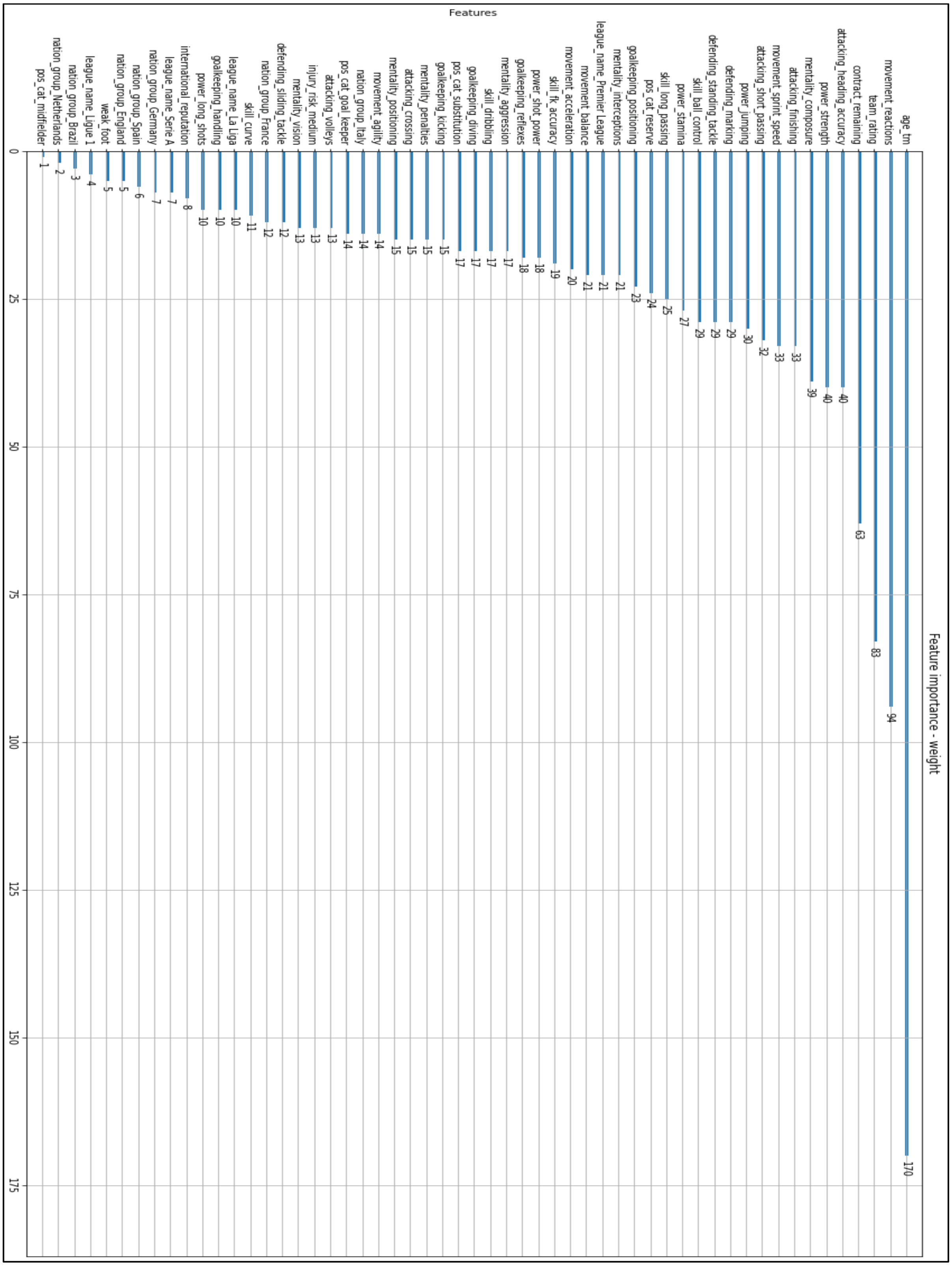


Figure A-1: Feature Importance-the Weight Metric

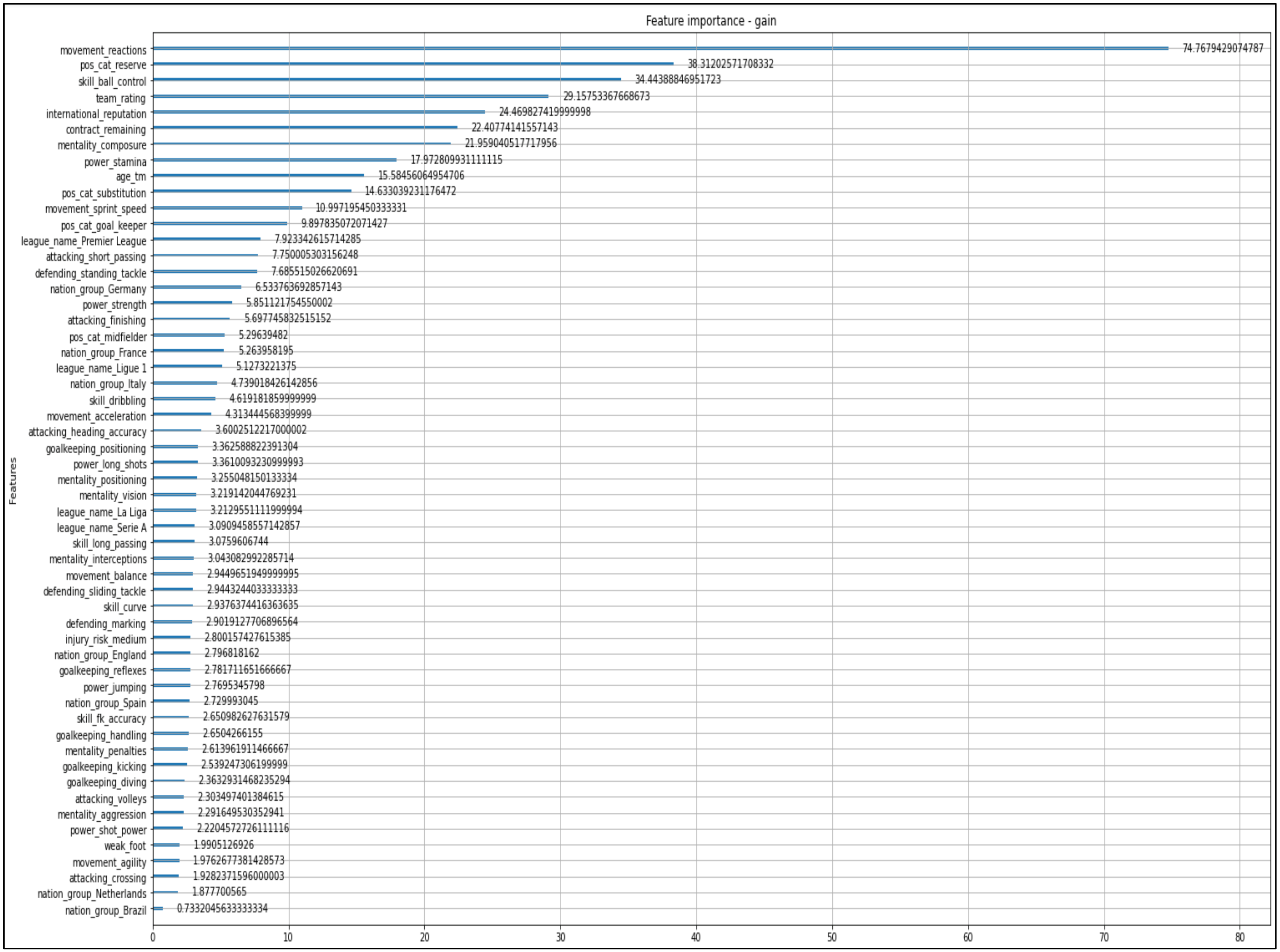


Figure A-2: Feature Importance-the Gain Metric

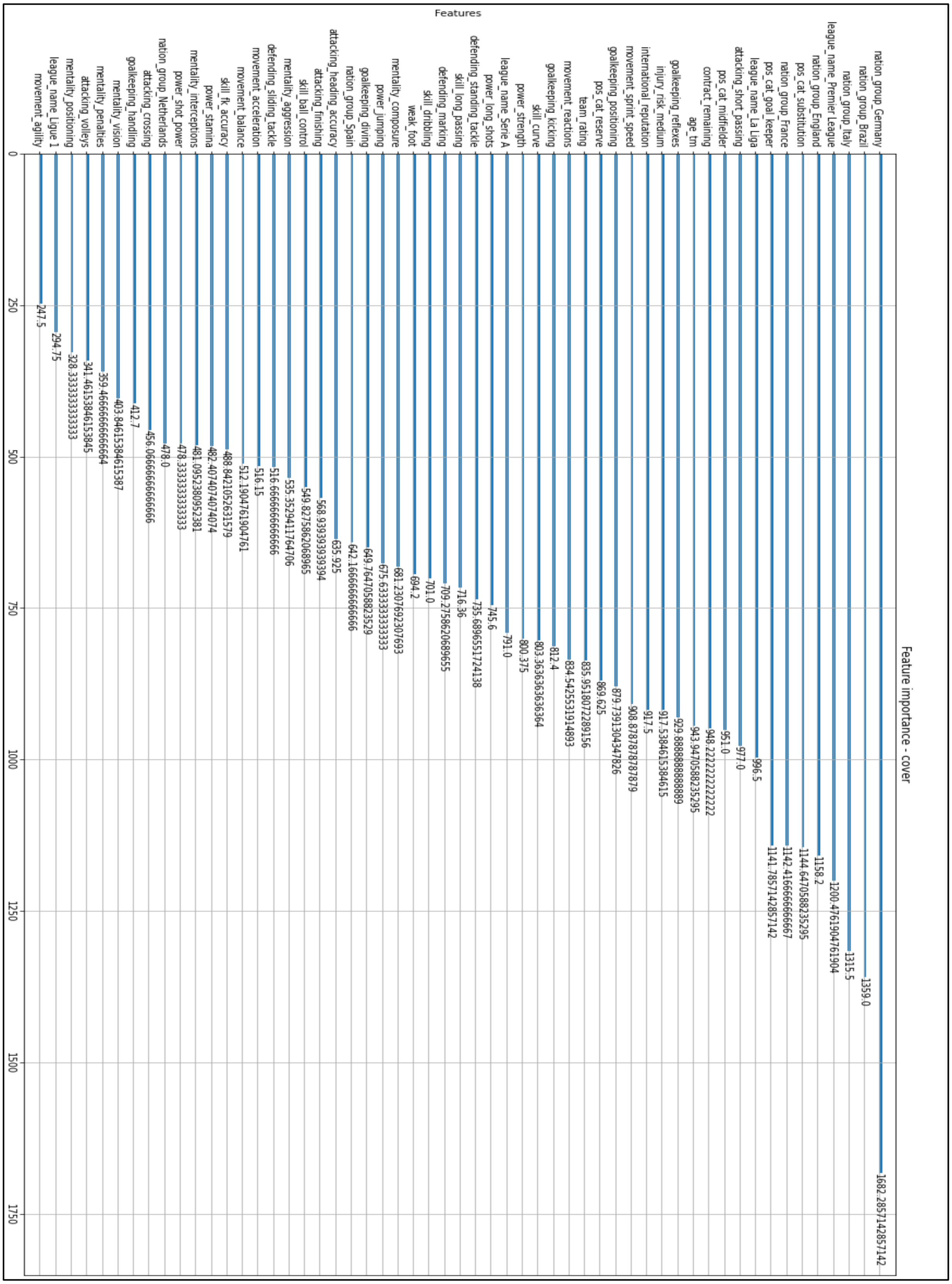


Figure A-3: Feature Importance-the Cover Metric