

A DATA DRIVEN PERFORMANCE EVALUATION FRAMEWORK FOR
SPORTS ANALYTICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYŞE ELVAN AYDEMİR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2021

A DATA DRIVEN PERFORMANCE EVALUATION FRAMEWORK FOR SPORTS ANALYTICS

submitted by **AYŞE ELVAN AYDEMİR** in partial fulfillment of the requirements for
the degree of **Doctor of Philosophy in Information Systems Department, Middle
East Technical University** by,

Date: 08.09.2021



I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ayşe Elvan Aydemir

Signature :

ABSTRACT

A DATA DRIVEN PERFORMANCE EVALUATION FRAMEWORK FOR SPORTS ANALYTICS

Aydemir, Ayşe Elvan

Ph.D., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Tuğba TAŞKAYA TEMİZEL

Co-Supervisor: Prof. Dr. Alptekin TEMİZEL

September 2021, 119 pages

Performance evaluation is a challenging, multidimensional and multi-criteria assessment problem. One application area is the player transfers in football (soccer), where player performance must be evaluated in-line with their responsibilities on the field. In this area of study, raw player performance statistics are not representative because of the external factors impacting the performance such as time-played, injuries, competition difficulty and characteristics, strength of the opponent, impact of actions in the game as well as the positions played. In addition, transfer market has unique financial dynamics in terms of transfer fees and player valuation. Some of the factors that affect transfer fees are athletic performance, properties of clubs and competitions and player popularity. The rich set of factors makes modelling transfer fees a challenging machine learning problem. This thesis provides a dynamic, context-dependent, probabilistic and hierarchical bottom-up approach for evaluating performance under uncertainty for custom requirements. Furthermore, the proposed framework links the performance metrics and various data sources to model transfer fees using machine learning ensembling methods. The proposed framework is generic and it can be adapted to other team sports.

Keywords: Sports Analytics, Unsupervised Learning, Player Performance Ranking, Player Valuation, Gradient Boosting

ÖZ

SPOR ANALİTİĞİ İÇİN VERİ GÜDÜMLÜ PERFORMANS ANALİZ ÇERÇEVESİ

Aydemir, Ayşe Elvan

Doktora, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Doç. Dr. Tuğba TAŞKAYA TEMİZEL

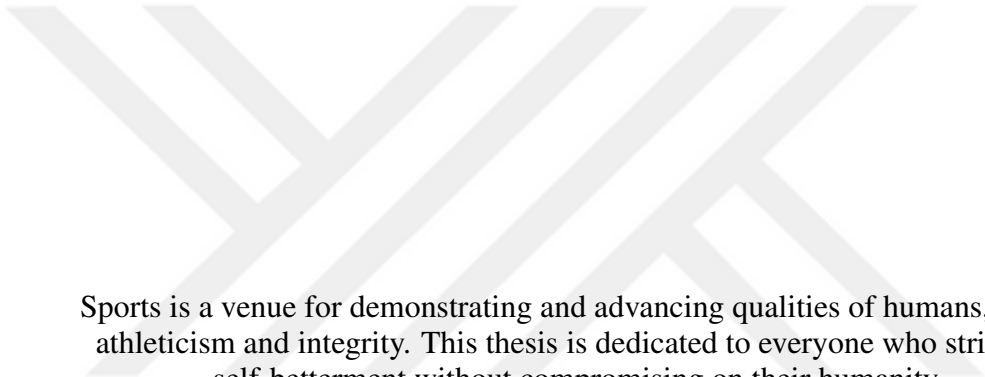
Ortak Tez Yöneticisi: Prof. Dr. Alptekin TEMİZEL

Eylül 2021, 119 sayfa

Performans değerlendirmesi çok boyutlu ve çok kriterli olmasının yanı sıra zorlu bir değerlendirme problemidir. Bu problemin uygulama alanlarından birisi olan futbolda oyuncu performansları oyuncuların sahadaki görevlerini ve istatistiklerini etkileyen diğer faktörleri (oyun süresi, rakibin gücü, müsabakanın stili) göz önüne alarak birçok değişken üzerinden değerlendirilmelidir. Bu alanda, oyuncuların saf istatistikleri oyuncu performansına yönelik bilgi verme konusunda yetersiz kalmaktadır. Bunun sebebi oyuncunun pozisyonunun yansıması, bu istatistikleri etkileyen birçok dış etmen bulunmasıdır. Bu etmenlerden bazıları, oyun süresi, sakatlıklar, müsabaka zorluğu ve karakteristikleri, rakibin gücü, oyundaki aksiyonların önemi, oyuncu pozisyonu vs. gibi faktörlerdir. Buna ek olarak futbolda oyuncu transferleri transfer ücretleri ve oyuncu değerlemesi bakımından genelden farklı finansal dinamiklere sahiptir. Oyuncuların transfer ücretleri oyuncu performansına ek olarak takım ve müsabaka özellikleri ile oyuncunun popülaritesi gibi faktörler bu dinamikleri etkileyen etkenlerden bazılarıdır. Bu çoklu etken seti, transfer ücretlerinin tespitini zorlayıcı bir makine öğrenmesi problemine dönüştürmektedir. Bu tez dinamik, olasılıksal, bağlama bağlı, hiyerarşik ve tabandan-tepeye bir yaklaşımla bilinmezlik altında performans analizini veri-tabanlı, gereksinim değişimlerine uyum sağlayan uyarlanabilir bir çerçeveye oturtmaktadır. Buna ek olarak, sunulan çerçeve performans değerlendirmeleri ve farklı veri kaynaklarını birleştirerek transfer ücretlerini makine öğrenmesi topluluk yöntemleriyle modellemektedir. Önerilen çerçeve genellenebilir olup diğer takım sporlarına da uyarlanabilir niteliktedir.

Anahtar Kelimeler: Spor Analitiđi, Denetimsiz Öğrenme, Oyuncu Performans Sıralaması, Oyuncu Deđerleme, Gradyan Arttırma





Sports is a venue for demonstrating and advancing qualities of humans, both in athleticism and integrity. This thesis is dedicated to everyone who strives for self-betterment without compromising on their humanity.

ACKNOWLEDGMENTS

I would like to express my gratitude, as much as I can, to my advisors Tuğba Taşkaya Temizel and Alptekin Temizel. I couldn't have asked for better advisors who have not only dedicated themselves to this research as much as myself, but also provided an unwavering support infrastructure for me as an individual.

I would like to thank everyone whom I had the pleasure of working with. The list is long however, I want to especially thank Laura Murphy and Anna Thomas for supporting me and my PhD studies with a sabbatical, and all my colleagues in Amplify Analytix who are amongst the best people I have ever had the pleasure of meeting.

I also would like to acknowledge two incredible women who have been my support system in the last few years, Sani Gaydarska and Alexandra Shkodrova. There were times I don't think I would have survived without your love, understanding and support. I love you both beyond measure and this thesis would not be complete without either of you.

I would like to thank my colleagues Kliment Preshlenov and Kris Nikolov for providing professional and logistic support, and taking over my work load at times I had to step away from work to focus on my thesis. Thank you for giving me the space to complete my studies.

To round up the colleague acknowledgements, I would like to especially thank Daniel Strahinov for being a complete rock to rely on both as a business partner and a friend.

Lastly, but most importantly, I want to thank my parents and my brother. I don't have the words for thanking you in any language but I would like to say that you are the most important things in my life and thank you for always being in my corner, regardless of what I choose.

TABLE OF CONTENTS

ABSTRACT	iv
ÖZ	v
DEDICATION	vii
ACKNOWLEDGMENTS	viii
TABLE OF CONTENTS	ix
LIST OF TABLES	xiv
LIST OF FIGURES	xvi
LIST OF ALGORITHMS	xviii
LIST OF ACRONYMS AND ABBREVIATIONS	xix
CHAPTERS	
1 INTRODUCTION	1
1.1 Background of the Problem	1
1.2 Statement of the Problem	3
1.2.1 Player Performance Evaluation	4
1.2.2 Player Valuation	5
1.3 Researcher's Motivation and Significance of the Study	6

1.4	Research Objective and Research Questions	8
1.5	Assumptions and Limitations	9
1.6	Thesis Outline	10
2	LITERATURE REVIEW	13
2.1	Data Collection	13
2.2	Sportive Performance Evaluation	16
2.2.1	Player Sportive Performance	16
2.2.1.1	Ranking Methods	18
2.2.2	Team Sportive Performance	19
2.2.3	Metric Development	19
2.3	Economic and Financial Modelling	20
2.3.1	Player Valuation	20
2.3.2	Financial Modelling for Clubs	22
2.4	Applications of Computer Vision	22
3	RESEARCH DESIGN	23
3.1	Introduction	23
3.2	Research Motivation and Research Questions	23
3.3	Player Performance	24
3.3.1	Context Representation	26
3.3.1.1	Scaling for Play Time	26
3.3.1.2	Scaling for Game Difficulty	32
3.3.1.3	Scaling for Competition Quality	33

3.3.1.4	Recency Scaling	34
3.3.1.5	Combining The Scales	34
3.3.2	Statistical Aggregation	35
3.3.3	Ranking	36
3.3.3.1	Preference Based Ranking	36
3.3.3.2	Rankings Towards an Objective	37
3.3.3.3	Rankings with No Feedback	38
3.4	Player Valuation	38
3.4.1	Player Value and Market Dynamics	39
3.4.2	Treatment of Transfer Fees	40
3.4.2.1	Continuous Transformations	40
3.4.2.2	Discrete Transformations	41
3.4.3	Factors Considered and Feature Engineering	42
3.4.3.1	Team Historical Transfers	42
3.4.3.2	Player Historical Transfers	43
3.4.3.3	Player Injuries	43
3.4.3.4	Player Popularity	43
3.4.3.5	Team Quality	44
3.4.3.6	Competition Quality	44
3.4.3.7	Team and Player Appearances	44
3.4.3.8	Player Role	44
3.4.4	Modelling	46

3.4.4.1	Model Training and Evaluation	46
3.4.4.2	Hyperparameter Optimisation	48
3.4.5	Model Ensembling and Information Extraction	49
4	EXPERIMENTAL SETUP	51
4.1	Player Performance Evaluation	51
4.1.1	Dataset	51
4.1.2	Parameters	53
4.1.3	Models	53
4.1.4	Metrics	54
4.2	Player Valuation	55
4.2.1	Dataset	55
4.2.1.1	Popularity Data Collection	56
4.2.2	Parameters	58
4.2.3	Models	59
4.2.4	Metrics	60
5	RESULTS AND FINDINGS	63
5.1	Player Performance Evaluation	63
5.1.1	Financial Validation	65
5.1.2	Team-Fit Validation	70
5.1.3	Information Retrieval Metrics	71
5.1.4	Real-World Results	73
5.2	Player Valuation	76

5.2.1	Discrete Models	76
5.2.2	Continuous Models	78
5.2.3	Stacked Models	81
5.2.4	Comparison With Existing Literature	81
6	CONCLUSION	87
6.1	Contributions	88
6.2	Limitations	89
6.3	Future Research Directions	90
APPENDICES		
A	PLAYER ROLE DEFINITIONS	101
B	FEATURE DEFINITIONS	105
C	MISSINGNESS STATISTICS	119

LIST OF TABLES

Table 1	Football Revenue Sources and Revenues for 2019/2020 Season	2
Table 2	Data Sources	15
Table 3	Definitions and Variables	29
Table 4	Major Factors Impacting Player Transfer Fees	42
Table 5	Dataset Restrictions	52
Table 6	Analysis Dimensions	52
Table 7	Parameter Details	53
Table 8	Parameters for Objective-Driven Rankings	54
Table 9	Ranking Setup for Player Valuation	58
Table 10	Hyperparameter Ranges	59
Table 11	Correlation Coefficients for Ranks Produced by Methodology	65
Table 12	Two-Sample Kolmogorov-Smirnov Test P-Values	68
Table 13	Rank Group Statistics Actual vs. Randomized	69
Table 14	The Results of the Proposed Solution using Information Retrieval Metrics	72
Table 15	Top 20 Players and Their Metadata	75
Table 16	Discretisation Statistics	76
Table 17	Classification Cross-Validation Metrics Across Years	77
Table 18	Overall Classification Metrics	78
Table 19	Balanced Accuracy by Class per Model	78
Table 20	F1-Score by Class per Model	78
Table 21	Learner and TM Performance in Transfers Between 8.5 - 50 Million €(RMSE in Million, N = 300)	80
Table 22	Prediction Performance on Test Set (RMSE in Million €)	80

Table 23	Meta-Learner Error Distribution on Test Set	81
Table 24	Top-10 Closest Predictions Above 500,000 €	82
Table 25	Bottom-10 Worst Predictions Above 500,000 €	82
Table 26	POS Initial Parameter Values	85
Table 27	Comparison to Existing Literature	85



LIST OF FIGURES

Figure 1	Frequency of Duels	27
Figure 2	Success Rate of Duels	27
Figure 3	Flow Diagram of Ranking Methodology	28
Figure 4	Frequency of Duels over Log-Transformed Play Duration	31
Figure 5	Success Rate of Duels over Log-Transformed Play Duration	31
Figure 6	Proposed Scale vs. Established Per 90 Estimations	32
Figure 7	UEFA Coefficients vs. Player Market Values	33
Figure 8	Distribution of Transfer Fees	39
Figure 9	Effect of Root and Log Transformations	41
Figure 10	Roles in Football	45
Figure 11	Multilabel Encoding	47
Figure 12	Sample Rows from Dataset	55
Figure 13	Search Term Frequency for Ronaldo	56
Figure 14	Search Term Frequency for Ronaldo and Football in Conjunction	57
Figure 15	Search Term Frequency for Burak Yılmaz and Forward in Con- junction	57
Figure 16	Search Term Frequency for Burak Yılmaz and Forward in Con- junction	58
Figure 17	Cross-Validation Folds	59
Figure 18	Cumulative Return on Investment per Age Group and Rank	66
Figure 19	Player Market Value Distribution per Rank Group and Analysis Period	67
Figure 20	Relationship Between Ranks and Post-Analysis Destination Team Ratings	70
Figure 21	Market Value Changes Through Time	74

Figure 22	Regression Performance In CV Folds	79
Figure 23	Error Distributions on Test Set Per Transfer Fee Buckets	83
Figure 24	Most Important 20 Features.....	84



LIST OF ALGORITHMS

ALGORITHMS

Algorithm 1	Weight Optimisation Procedure	37
-------------	-------------------------------------	----



LIST OF ACRONYMS AND ABBREVIATIONS

AFC	Asian Football Confederation
CAF	Confederation of African Football
CAM	Central Attacking Midfielder
CB	Center Back
CDM	Central Defensive Midfielder
COMET	Characteristic Objects METHod
Concacaf	The Confederation of North, Central America and Caribbean Association Football
CONMEBOL	The South American Football Confederation
CPP	Comparison of Probabilistic Preferences
EPL	English Premier League
FB	Full-Back
FIFA	Fédération Internationale de Football Association
MCDM	Multi-Criteria Decision Making
OFC	Oceania Football Confederation
UEFA	The Union of European Football Associations
WB	Wingbacker



CHAPTER 1

INTRODUCTION

With the increased amount of data, advanced analytics and machine learning has been transforming most industries from healthcare to manufacturing and the sports industry is no exception. In fact, the sports industry has been using analytics for decades for betting purposes. However, with the development of data extraction and gathering technologies such as computer vision and wearable devices, the number of application areas for analytics had a tremendous increase, which is evident from the number of recent publications in the field [1].

While accurate as an umbrella term, "sports analytics" glosses over the fact that it is an independent and emerging area of complex research. For instance, analytics for team sports and individual sports would vastly differ from each other, so would analytics on a sensor level (i.e. wearables) to competition level (i.e. aggregates of statistics for several competitors). In addition, there are significant differences between sports disciplines and their rules. For example, while they are both sports, analytics for chess would vastly differ from analytics for lacrosse. There are 67 sports disciplines included in Olympics 2020, which is a mere fraction of all sports disciplines practised in the world. Given the breadth and the depth of this emerging field, we expect research efforts to accelerate rapidly.

Out of all the sports, Association Football (Football from now on) is the undisputed biggest spectator sport of all time [2]. This thesis focuses on building an analytical framework for Football to assess player performance and bridge the gap between performance and financial valuation. The proposed framework is designed to be applied to other team sports, on a global scale as opposed to the existing approaches in the literature which work with a limited scope in terms of region or competition.

1.1 Background of the Problem

Football clubs regularly compete in their domestic leagues and domestic cup competitions. Best performing clubs in these domestic competitions compete in the international leagues and cups, depending on the association they are part of. For instance, most European countries are part of The Union of European Football Associations (UEFA), and the best performing clubs in these countries get to compete in Europa League or (European) Champions League. There are 6 confederations, which are

Table 1: Football Revenue Sources and Revenues for 2019/2020 Season

Revenue Source	Revenue Amount (in bil. €)
Media Rights	25.9
Commercial Rights	4.2
Ticket Sales	0.38
Other (incl. Transfers)	23.3

collectively governed by FIFA, in the world: AFC, CAF, Concacaf, CONMEBOL, OFC, UEFA. The clubs from these confederations compete in world cups every 4 years, which provides the biggest viewership and exposure to players and the clubs [2].

Club revenues come from various sources. Media and commercial rights are the most significant ones which are driven by domestic and international performance. Transfers are also a major revenue source for clubs. [2]. Total revenues incurred from various sources are given in Table 1, which is adapted from [2]. In 2019, the size of the player transfer market for European clubs was 28.9 billion € [3]. Due to the global pandemic, in 2020, the revenues from all sources have decreased, however with Qatar 2022 World Cup on the horizon, FIFA is optimistic the finances will recover.

Clubs have a specific transfer strategy to incur revenue and invest in players for increased performance [4]. As commercial entities, clubs are required to balance their revenue and spending, all the while optimizing their club roster for performance and various objectives. In fact, the club finances are governed by a set of rules, called "Financial Fair Play" [5] which restricts club expenses on players and additional expenses, if their balance sheet is out of order.

Player transfers are a unique item in financial dynamics of football clubs. Transfers are a tool for revenue generation and performance adjustment. However, they are also highly dependent on the athletic performance of the individual players [6]. Therefore, on one hand the acquisition and release of players behave exactly like asset management. On the other hand, the asset in question has the individual ability and will to self-develop or otherwise. In addition to the athletic development of an individual player, the personal brand is also an important factor for developing the player as a financial asset. Both athletic performance and branding are affected by the environment and demographics of the player [7], [8].

Furthermore, the transfer market has shown to be very inefficient in pricing [9]. Very few among all players, so-called super-stars, are valued significantly much higher than the rest of the player population [7], [10]. In fact, it has been shown to follow a power-law distribution which does not allow for traditional economical analyses and calls for more complex methods [11], [12]. This discrepancy is due to impact of factors other than performance on values of players, which makes analysis and modelling a complex and challenging task.

In the light of these facts, it is trivial to see that the football clubs have an incentive to perform at their best while optimizing their finances jointly with their performance. Given the size of the returns, the competition is fierce and even the modest optimizations are beneficial and desirable for the clubs.

Intuitively, performance optimization and financial optimization can be conflicting objectives [8]. Simply put, building a roster from the best players in the world, although uncertain, may result in the best performing team (ignoring the factor of player synergies). And yet, this would be very costly and only a few clubs, if any, in the world would be able to afford such an expense. Instead, clubs must evaluate the players based on their athletic performance requirements and make transfer decisions based on their financials. This is a complex business optimization problem because there are over 100000 active football players in the world, and yet clubs only have access to (or even knowledge of) a small subset of these players.

Traditionally, the scouting process, i.e. the process of discovering and signing players, was done in person where scouts of a club would personally go and watch players and decide whether they would be a good fit [13]. In addition to being costly, from a time and money perspective, this kind of process also limits the available pool of players to a handful of players scouts could effectively observe. In addition to this, cognitive biases in human decision making and other personal biases make this process highly subjective and inefficient. Recently, and even prior to the pandemic, this process has been assisted by watching player videos that only show highlights of player actions as opposed to the full performance in each game. In addition, it is infeasible for a scout to watch all games of a player, whether in person or through video, to make an informed decision.

Recently with the advances in data collection for football, this process has begun to change from these inefficient methods to more data-driven approaches [13]. However, the current state of the art in data-driven scouting is limited to analyzing competitions or players in isolation, therefore it is still a limiting factor. The inability to analyze and compare players from different competitions and regions makes the decision process similarly inefficient and limited and leaves room for human bias.

1.2 Statement of the Problem

This thesis aims to build a comprehensive framework to perform player athletic performance evaluation and tie it to the player's financial value in the transfer market in a data-driven way. There are various factors that affect both player performance and player's transfer value (or transfer fee) **Quantifying Performance Individuala**, [14]. These factors must be defined in a fashion that is extendable to other sports and quantified using the various data sources that are not linked together, to arrive at a globally applicable and comprehensive team sports player assessment methodology.

The problem consists of two sub-problems: player performance evaluation and player value modelling. The following sub-sections state these problems in formal terms.

1.2.1 Player Performance Evaluation

Traditionally, in-game statistics are used to evaluate player performance. These statistics usually represent frequencies of actions such as the number of shots taken by a player and success rates (i.e. percentages) such as the percentage of shots on target over all shots. While these statistics are interesting from a spectator standpoint during the game, they provide very little information on their own for comparison of different players across games. To address this, analysts typically use per-90 averages across multiple games [15]. This approach allows a relative comparison of players in the same competition, however, it is still limited in terms of providing a useful measure for player performance that can be applied to everyone because player performance is affected by various contextual factors. It is also interpreted differently under different contexts.

Consider a scenario where a comparison of two players is required. *Player A* plays in *English Premier League (EPL)* which is one of the highest-quality competitions, if not the highest overall. *Player A* is a highly experienced star player who typically gets to perform for the full duration of the match. Conversely, *Player B* is an up and coming young talent in *Turkish Super League*, which, while being a solid competition, does not compare to *EPL* in terms of quality. *Player B* is a young talent, therefore has a lower duration play-time. Assuming these players are playing in the forward position, it is plausible that these two players can have the same number of shots taken in a game (despite the difference in their play-times) and have the same success rates. However, intuitively, we would evaluate these players differently due to latent factors. Assuming both players consistently have high statistics, we would be likely to ascertain *Player A* as a world-class player in their role, while *Player B* would be deemed a highly-promising young talent but with an unknown estimation of potential, therefore would not be rated as highly.

Therefore the problem of comparing all active athletes in a team-sports can be defined as in Definition 1.

Definition 1. Arrive at a global ranking R , for player's role or similar roles POS , given a set of in-game statistical measurements M , a set of contextual factors C that define the conditions in which the match is played and the player is performing.

As stated in Section 1, clubs may be looking to optimize different objectives. These objectives are often tied to the performance of the players they have or are looking to acquire. In such cases, the task of player performance evaluation can be further clarified as in Definition 2.

Definition 2. Arrive at a global ranking R , for player's role or similar roles POS , given a set of in-game statistical measurements M , a set of contextual factors C that define the conditions in which the match is played and the player is performing such that R will maximize ranker objective OBJ where statistical performances of two or more players are comparable.

Despite the fact that the roles in Football are defined in terms of the responsibilities of the player (for instance, we expect good forward players to take accurate shots and score goals), clubs may value certain characteristics of a role more than others. For example, clubs may look for a secondary forward player to draw attention from the opponent's defenders while the primary forward would be expected to score goals. In such cases, the player responsibility (or dimension) of scoring a goal would be less important than their dribbling ability. In such cases where clubs or scouting entities have already defined specific requirements in which they want the player to perform well, these preferences must directly be used to arrive at custom player rankings. Therefore, in these cases, the player performance evaluation problem is defined in Definition 3.

Definition 3. Arrive at a global ranking R , for player's role or similar roles POS , given a set of in-game statistical measurements M , a set of contextual factors C that define the conditions in which the match is played and the ranker preferences W .

To summarize, this thesis proposes a novel framework to rank players by introducing the following contributions to the literature:

1. A simple, yet powerful context embedding method to align the in-game player statistics obtained under different contexts on the same scale. This method shall further be named "Scaling".
2. An extendable set of enriched statistical aggregations to quantify various performance dimensions depending on the player position.
3. A flexible player ranking approach to address various ranking scenarios as specified in Definitions 1, 2 and 3.

1.2.2 Player Valuation

Player valuation is the second major component of the proposed framework. It is a complex estimation problem with various aspects due to the complex structure of the transfer market.

There are several types of player transactions in the transfer market. The first one, which is the focus of this thesis, is the paid transfers. In this type of transfer, clubs pay a transfer fee to the player's current contracted club as a release clause. This shows that the buying club has valued the player to pay a certain fee, and this is the clearest indication of a player's value. Other than paid transfers, a player may also be released from their club without a fee. Usually, such free transfers happen when a player's contract expires or the club has some other incentive to waive the fee. In addition, clubs can loan players to other clubs either with or without a fee. Finally, there are "promotional" transfers in which youth players get promoted to the parent club.

Player transfers in football is a major revenue source [3]. Yet despite the size of the market, it is reported that the clubs are inefficient in transfers [6], [16], often making sub-optimal transfer decisions, resulting in unsatisfactory player athletic performance or eventual financial gain when selling the player. Therefore, it is important for clubs to know the fair actual market value of the player given various factors.

The existing studies in the literature that model the player values in football focus on understanding these factors that impact the player values [17]–[19]. One such factor is the properties of the buying club. However, the player’s contracted club has no way of knowing which club would end up buying the player before transfer negotiations start. Therefore using buying club properties in the modelling process makes existing models inapplicable in real life, albeit they are useful as explanatory models.

Player values are also dynamic estimations, changing based on their recent performance or age. Player values also change based on their tenure in the contract and players are more likely to be part of free transfers if they are reaching the end of their contracts. Therefore any model that predicts player values must be built on the latest available information with continuously updated performance measurements and latest metadata for players, like injuries. Furthermore, as shown in [7], [10] valuation of famous players (i.e. superstars) is exponentially higher than a randomly selected player. This super-star effect must be included in the modelling to reflect public opinion. Like all the other factors that impact player values, public opinion is also dynamic (i.e. time-dependent), therefore must be modelled as such.

Given this context, this thesis frames the problem of player valuation as in Definition 4.

Definition 4. Build a dynamic predictive model using player performance rankings and measurements in Definition 1, player demographics, information regarding player’s club before transfer, injuries and player popularity in public opinion to estimate player’s fee in case of an immediate paid transfer involving the player.

1.3 Researcher’s Motivation and Significance of the Study

This study is motivated by the existing gap between data availability in Football and utilisation of the available data. In recent years, data availability increased significantly with many data providers providing more data, both in volume and depth. Yet, the analytical approaches in football remain relatively simple. Given that football industry is home to many high-value transactions and the environment is highly competitive for a large amount of finances, addressing and closing this gap is of interest to stakeholders in the football industry, especially clubs and representatives of players. This gap is already being acknowledged by clubs and even FIFA, which is moving from simple analytical approaches to more predictive methods to fairly evaluate player values for financial fair play purposes.

In addition to the contributions in terms of its utility in the football industry, this thesis also aims to provide a mathematical foundation for the emerging field of sports

analytics. Unlike the existing literature, we aim to approach the scoped problems from a broader perspective and provide means of developing progressive applications on top of the proposed framework. This maturing field currently focuses on small-scale analyses and applications, we aim to provide a scalable and broader framework.

Furthermore, the existing approaches study athletic performance and financial problems separately. However, in the context of team sports, these problems are tightly coupled. Previous studies have shown links between player salaries and player performance, however, no specific effort has been made to jointly study the two sides of this coin.

To simply put, this thesis aims to address the following gaps in the literature:

- Player performance analysis is done for small local subsets of all available players. This makes comparison across different subsets impossible, limiting the pool of players clubs or other stakeholders can make a data-driven decision on.
- Player ranks are dependent on rankers' requirements or objectives. However existing ranking methods do not allow for different rankers' inputs. Instead, they are based on "expert opinion", which makes them unreproducible and subjective.
- Player performance is highly dependent on latent factors which are not taken into account in the existing literature.
- Existing player valuation models are either exploratory, or they use synthetic data from sources such as FIFA or Football Manager video games. Such a dynamic problem must be based on real-world data sets and must be addressed using predictive models capable of modelling complex datasets and interdependent predictor variables with incomplete information.
- Player valuation is a complex problem that requires combination of player roles, demographics, club and player performances, social media and player popularity metrics and player performance rankings.

To address these gaps, this thesis provides the following contributions to the literature:

- First globally applicable data-driven player ranking methodology.
- A new ranking framework that is extendable and has the ability to incorporate specific requirements. This framework can be applied to other team sports, for which performance data is available, as well as can be extended within the scope of Association Football.
- A data-driven "baseline" ranking in cases where expert opinion is unavailable. Existing methods rely on definitions and criteria provided by experts. We circumvent this problem by using unsupervised learning techniques to arrive at an overall ranking.

- An optimization scheme to find the combination of skill-sets in a player to achieve an objective. This scheme is applied to discover the player performance profiles that maximize the return on investment, however, it can also be applied to identify the profiles that contribute most to club placements or win likelihood. In other words, to identify key skills for sportive success.
- A "customisation" layer to arrive at player rankings that satisfy specific and well-defined requirements. This layer utilizes distance algorithms to reflect the ranker's criteria to the rest of the proposed solution.
- A "layered" approach to the overall ranking problem which builds hierarchically from raw in-game statistics to final player ranks. This hierarchical algorithm allows for customisation and adaptation at each layer to different use-cases and sports.
- A data-driven player transfer fee estimation model that combines player roles, demographics, club and player performances, social media and player popularity metrics and player ranks that uses advanced machine learning techniques.
- Application of advanced model ensembling techniques to combine predictions from various fee estimations to optimize the final predictions to different scenarios. The complexity and the dynamism of the transfer market are addressed by using multiple predictive models and their intelligent and adaptive combination.

1.4 Research Objective and Research Questions

The research objective of this thesis is to systematically approach the problem of player evaluation and valuation in team sports. I aim to provide a solid analytical and mathematical foundation for the analysis of the scoped problems. We also aim to link two seemingly separate fields of study in football analytics using advanced analytical approaches.

In addition to accelerating the adoption and application of machine learning in a new field of study, we also aim to use this complex field to study the behaviour of various machine learning approaches in application. Specifically for estimating player transfer fees, predictive machine learning models are a novelty. There is no established best practice in representation of this problem using such techniques with real data. Capitalizing on this gap, I use this opportunity to compare and contrast different representations of the target variable and aim to introduce new ideas in modelling similar problems.

The research questions of this thesis, therefore, are as follows:

- Player Performance Evaluation
 - Player performance evaluation is limited to basic in-game statistics that are aggregated in isolation. Comparison between different competitions

and regions is impossible. Is there a data-driven way of comparing players?

- This comparison requires a systematic way of including external factors normally considered by experts to be included in the calculations consistently. Is there a mathematical representation that could allow external factors to be added?
- Are there additional statistical methods that go beyond simple aggregation to evaluate player performance?
- Rankings depend on the criteria of the ranker. Is there a mathematical way of including ranker preferences?
- Is there a robust way of getting rankings where preferences are unknown or cannot be quantified?

- **Player Valuation**

- What are the major factors that impact a player's transfer fee?
- Football transfer market is very dynamic. How can machine learning models help with estimating the fee of a transfer?
- Player values depend on media hype, fan expectations and connectedness of the player. How can this information be added to the more standard data such as player age or other demographics?
- How do player ranks relate to player transfer fees?
- How can different predictions and estimations be combined into a more robust and accurate transfer fee prediction?
- What is the performance and applicability of different modelling representations on a skewed and dynamic outcome variable?

1.5 Assumptions and Limitations

This thesis relies on data collected from live football games by data providers, as well as additional data sources available online. The assumption in this thesis, as in any data-driven study which uses real-world datasets, is that the data collected by individual operators and data available online is of suitable quality and internally consistent.

The main assumption made by this thesis for the player ranking module is that the contextual factors can be encoded using formal mathematical representations. The contextual factors must initially be identified and defined, then they must be represented mathematically to enrich the in-game statistics. One limitation of this approach is that there may be multiple ways to represent similar context by different researchers, and the comprehensiveness of the context embedding is dependent on the design by researchers. However, the aim of this thesis is to provide a generally applicable framework and not to assert a certain set of mathematical representations,

therefore we hope to initiate further research by providing an initial set of representations.

This study focuses on modelling paid transfers. To formally define the scope of player valuation, this thesis models the player transfer fees assuming an immediate paid transfer for any given player. However, paid transfers only make about 15% of all transfers in the transfer market. The remainder of the player movement in the transfer market is either loans or free transfers. While this thesis does not analyse these different types, they are certainly of interest for this field of study.

Predictive approach to player valuation component comes with certain trade-offs. In order to use predictive modelling, any information post-transfer of the player had to be discarded. Studies have shown that properties of the buying club is a major factor in determining the transfer fee, however, to avoid so-called "data leakage", any information post-transfer is omitted from the dataset which results in a trade-off between accuracy and real-world applicability.

Finally, the "connectedness" of the player relies on knowing their connections in the transfer market. Football players are typically represented by another person or an entity (player agents) that handles their dealing with prospective and current clubs. They not only serve as a representative for the player, but they can also serve as a marketing agent. Therefore the connections of this representative can be a major factor in promoting the player to clubs that are in a scouting process. Unfortunately, while the current agents of players can be found on Transfermarkt, historical agent data is not available. Therefore this research question must be answered in the future when data such data is available. To partially answer this question, this thesis quantifies player popularity.

1.6 Thesis Outline

Chapter 2 provides an extensive literature review to analytics in the sports industry with a particular focus on football and studies directly relevant to this thesis.

Chapter 3 details the proposed framework and methodology by providing research design and mathematical formulations. This chapter is organized to have separate subsections for player performance evaluation and player valuation. It also contains specifics of research questions.

Chapter 4 provides the experimental setup for the application of the proposed methodology to a comprehensive, real-world dataset. Dataset properties are also provided and compared to the existing literature to illustrate the relative scope of the study.

Chapter 5 compares the results to existing literature where comparisons can be made. This is an emerging field, so the number of studies with real-world comparable datasets is fairly limited. However, we scanned the literature thoroughly and included the state-of-the art in comparisons.

Finally, Chapter 6 summarizes the findings and contributions of this thesis. It provides pointers for further research questions, as well as new application areas in terms of other team sports.





CHAPTER 2

LITERATURE REVIEW

Outside of betting analytics, the earliest known analyses regarding sports date back to the 1960s. An English pioneer, Charles Reep, paved the way for using data and analytics to guide English Football Club Brentford's in-game tactics [20]. His observations were such that, teams that have 4 or fewer passes towards the opponent goal scored more. Based on this statistical observation, he advised teams he was working with to perform long-ball plays in which clubs perform long passes towards the opponent goal. He influenced English football for decades, which led to controversy due to poor national team performance.

Charles Reep is much lesser known than Michael Lewis, the author of the now infamous book and movie "Moneyball: The Art of Winning an Unfair Game" [21]. Lewis, in his book, provides detailed approaches to player analytics, roster formation and success in the context of baseball. Applying his ideas, he proceeded to achieve great sporting success with a limited budget.

The "Moneyball" concept has been adopted by many practitioners in various sports [22]–[25]. This thesis is also heavily influenced by the ideas introduced in his seminal work, namely the idea of efficient allocation of resources under performance pressure.

The increase in data collection allowed for the field to develop further. The field has now branched out into different areas of specialization whose literature shall be reviewed in corresponding sections. This review mainly focuses on football-related literature however, where suitable, we also include literature from other team sports.

2.1 Data Collection

With the advancement of technology, data collection methods have also progressed. There are now several companies that perform semi-automatic data collection from sports broadcasts, as well as wearables [26]–[29]. In addition, there are also websites that crowd-source data regarding players, teams and transfers. In the context of football, the biggest such website is Transfermarkt [30]. Table 2 provides a summary of football-related data sources. In this table, "events" refers to individual actions taken by players such as passing the ball. This data type typically comes with a timestamp and several attributes such as whether the action was successful or not (i.e. pass re-

ceived by a team-mate) and sub-types of the event (e.g. whether a pass was vertical or long). In-game statistics is built up from these events into per match and per player action frequencies or success rates. The events are typically actions taken on the ball. In addition to these action statistics, data providers are soon aiming to provide positional data regardless of whether players are on the ball or not. We expect this next step to spark a new set of spatio-temporal analytics.





Table 2: Data Sources

Data Source	Coverage	Collection Method	Data Verification	Match Outcomes	In-Game tics	Statis- tics	Events	Transfers	Financial Info	Club Rank- ings	Country Rankings
TransferMarkt [30]	Major competitions and players	Crowdsourced	Expert Opinion					X	X		
UEFA [31]	UEFA Member Association	Expert Opinion	Expert Opinion							X	X
WhoScored [32]	Major competitions and players	Aggregated from Opta Sports	Unknown	X	X						
OPTA [28]	Over 1000 competitions and 200k players	Human experts using bespoke analysis software [33]	Expert Review	X	X		X				
WyScout [29]	Over 1000 competitions and 200k players	Human experts using bespoke analysis software [34]	Expert Review	X	X		X	X			
InStat [26]	Unknown	Unknown	Unknown	X	X		X				
StatsBomb [35]	Unknown, however the website claims more granularity than any other provider	Unknown	Unknown	X	X		X				

2.2 Sportive Performance Evaluation

Sportive performance evaluation refers to the set of problems in which the sporting ability is subject to measurement.

In team sports, a distinction must be made between players' athletic performances and teams' athletic performances. While teams consist of players, the sporting success of a team is dependent on more than the individual athletic capacity of teams' players. Therefore we divide this section into player and team athletic performance.

2.2.1 Player Sportive Performance

Player performance is one of the main fields of study in sports analytics. Starting with Michael Lewis [21], it became important to quantify the performance of the athletes given their responsibilities.

In [36], authors conduct expert reviews in Basketball to assess the quality of the existing player performance metrics. They report that the experts find the existing techniques of performance evaluation insufficient. In [25], authors model the match outcomes using linear regression techniques to identify independent player performance variables that affect the match outcome. Then, the authors propose recruiting players who perform well in these variables. This is one of the simplest methods in data-driven scouting. However, it does not explicitly analyze player performance.

In [37], authors build a probabilistic model that quantifies the impact of player actions on match outcome in Hockey games. They aggregate the action impact per player to arrive at player performance measures. Similarly for Hockey, in [38] authors apply Markov Decision Processes and K-Means clustering to group and quantify impact of actions to evaluate performance of the players.

In [39] authors work back from existing player ranks and validate the rankings using correlation analysis and partial decision trees with match outcomes. They aim to circumvent the subjectivity arising from expert rankings by validating them with data.

In [40], authors use Grey Relational Analysis to quantify the player performance based on match statistics in NBA.

In [41] authors take a network modelling approach to quantifying the impact of player actions in a match. They model the action sequence as a network and use flow centrality to estimate the impact of the actions on the match outcome.

In [42], authors build expert-guided heuristic linear regression models to rank players. They propose six separate indices to quantify player performance that additively combine into a final player performance index. They highlight that football is a highly complex game that cannot be fully represented as such. However, this research was driven by Football Association (FA) in England to rank players. This is the current official ranking methodology of FA.

Similarly in [23], player performance indicators are identified by experts per position. Players are then assessed for their performance on these indicators.

The action importance problem is narrowed down to quantification of the impact of passes in [43] and [44]. In [43] authors use supervised learning techniques to model and predict whether the pass would end in a shot or not. This limits the evaluation of performance only to offensive abilities. In [44], the passes are modelled on their likelihood of resulting in a shot or not, however in this case the authors include a 'risk' quantification to further refine the assessment of passes.

In [45] authors build a descriptive approach that can differentiate between key passes which have a positive impact on the match outcome and regular passes which are not as impactful using a computationally light approach.

In [46], authors employ reinforcement learning techniques to quantify the value of the actions of players. Out of all action valuation methods in the literature, this is the most advanced and nuanced one that does not make a Markov assumption. Instead, the reward function is learned from data to quantify actions.

In [13], authors aim to predict the transfers of players between teams given their match statistics. To simplify the problem, they group the movements as 'moving to upper tier', 'staying in the same tier' and 'moving to lower tier'. They show that player performance is predictive of transfer movements.

As a more recent example of expert analysis in football, in [47] authors perform a survey to identify the variables that are important to analyse the attacking performance and employ questionnaires to rate the importance of each individual variable. The weights obtained through expert opinion, pass analysis and descriptive statistics are then standardised to arrive at the player attacking index. The approach is applied to the players of Atletico Madrid football club.

The following three recent studies are of particular interest to this thesis. In [34] authors estimate the roles of the players using k-means and rank the players based on their actions within the said roles. They validate their approach based on expert opinion. They apply the proposed methodology to multiple European competitions of comparable quality.

In [1] authors use CPP-Tri ranking method to evaluate players and arrive at rankings. They validate their results based on the financial performance metrics of the player. They apply their proposed methodology to a limited number of players.

Finally, in [48] authors apply the COMET [49] method to rank 6 football players respectively. COMET is a novel multi-criteria decision making method that uses fuzzy logic.

The approaches listed so far are all limited in their scope and applicability. Expert surveys suffer from human limitations. There are only so many factors individuals can consider. Experts can also evaluate and validate a limited number of players.

Similarly, other modelling approaches struggle to differentiate between player performance when player statistics are the same but players come from different contexts.

Conversely, action quantification methods suffer from the opposite problem. The actions themselves are too granular to build a picture of fluidity, tactics and strategy. Simply accumulating actions do not necessarily reflect the larger picture all player actions jointly build up. This kind of "sum is bigger than the parts" situation is called an "emergent phenomenon" [50]. Furthermore, aggregating actions in different competitions may result in the same player performance index despite players could be vastly different in their quality.

These studies provide valuable insight to the important factors in player performance. They have also guided this study in terms of method evaluation methodologies.

2.2.1.1 Ranking Methods

Most methods that evaluate athlete performance arrive at final performance rankings. The most established tools for arriving at rankings are a set of algorithms called "Multi-Criteria Decision Making Methods" (MCDM). These methods model multiple sets of criteria in a probabilistic manner and try to place queries in a final order based on the decision making points and assumed statistical distribution. There are several methods that fall into this category. A comprehensive description of MCDM methods can be found in [51].

There are two studies that employ MCDM methods to the player ranking problem [1] and [48]. In [1], authors employ the more established Composition of Probabilistic Preferences (CPP) algorithm to rank the players. CPP methodology performs numerical estimations of preference probabilities through combinatorically comparing categories with alternatives, then calculating overall probabilities of an alternative being better than the rest. Being a combinatorial algorithm, CPP has a polynomial time complexity $O(n^k)$ where n is the number of alternatives and k is the complexity exponent ($k > 1$). Therefore, with the increased number of alternatives (i.e. number of players), computation of ranks takes significantly longer. CPP has multiple variants based on the underlying probability distribution assumption.

The second relevant MCDM method is COMET [49]. Unlike CPP, COMET does not rely on alternate combinations. Instead, it makes use of fuzzy decision criteria. The so-called "Characteristic Objects" are derived based on expert-driven decision making points for each variable. In addition, there is also a need for "expert decision function" to sort the memberships in characteristic objects. Based on each instance's characteristic object values, "Matrix of Expert Judgement" is created and then each instance is ranked based on the decision function. COMET does not suffer from the time-complexity problem of CPP, however, it relies heavily on expert design prior to execution of the method.

2.2.2 Team Sportive Performance

In most cases predicting the match outcome is synonymous with team sportive performance. In [52] authors adapt ELO ratings [53] into football, and use these ratings to predict match outcomes. As an extension to this work, in [54] authors use the ELO ratings of teams to arrive at bookmaker odds.

Instead of modelling the match outcome directly, in [55] authors model the goal differential of teams, and quantify the impact of transfers on team performance in terms of goal differential. They employ semi-Markov Decision Process modelling to evaluate team performance.

In [56], authors attempt to predict team ratings provided by official sources using supervised learning methods. First, they aggregate player in-game statistics to team aggregates, then they use the statistical dataset to predict official ratings. They also calculate ELO ratings and compare them to official ratings. They find that ELO ratings correlate better with official ratings and they also have low variance. This study is important as a negative result. It also demonstrates the difficulty of aggregating statistics in a meaningful way in the context of Football.

In [57], authors rank the clubs using COMET method and correlate their rankings with WhoScored.com [32] rankings.

In [58], authors combine match statistics with spatio-temporal video data to partially explain the so-called "home advantage" in which the home team is deemed to have a slight advantage over the away team. They demonstrate that home teams are able to take better control of the final-third of the pitch and create more chances. They hypothesize that this phenomenon partially occurs due to away teams playing more defensive.

To model the team tactics, two papers lead the way. In [59], authors make the assumption that teams follow a pre-defined number of tactics in a game. They model the player positions on the pitch using a probabilistic representation of all tactics per position. Then authors try to discover the distribution of tactics using LDA [60] methods. Conversely, in [61] authors build uninterrupted action sequences from events data to automatically discover clusters of events to determine the tactics.

2.2.3 Metric Development

Metric development is an interesting but not formally studied area of analytical application in the context of Football. In Economics and Econometrics, this field is very important. The Economic indicators are developed (GDP, CPI, Manufacturing PMI, Inflation Rates etc.) to reflect a certain performance aspect of the economy. There are similar, yet unstructured efforts in Football to quantify club or player sporting success using standardized metrics.

Instead, individual practitioners develop these metrics based on the need for analysis and interpretation and the adoption happens depending on the ease of calculation for the metric and the value of the information it provides.

The best adopted metric in sports analytics literature is "Expected Goals" [62]. There have been several improvements over this metric across the years, however, the main idea has remained the same: Quantify the number of expected goals a player or a team will score in a match. This metric allows for understanding whether the athletic performance of a player or a team in a given match is above or below expectations. One extreme example comes from Bundesliga. In the season of 2014-2015, one of the best teams in the competition, Borussia Dortmund was in the relegation zone despite playing decently. Based on their game-play, the expected goals were decent however the club failed to score close to what the metrics proposed. In the case of not having the guidance of this metric, clubs would be expected to make changes to the coach or the players however, Borussia Dortmund as a data-driven club were aware that the lack of goals in the team were simply due to uncontrollable factors and if they keep playing consistently, their scores would rise up to the expected goals. They finished the season on the 7th place, which is not excellent however far from relegation zone [63].

Ideally, the metrics developed would reflect an important piece of information given the object of analysis. A summary of relevant statistics per player and role can be found in [23]. This thesis makes use of the metrics listed here to quantify the performance of players in different roles, as well as additional metrics.

In addition to statistics relevant to the individual players, there are several metrics that quantify team behaviour on certain dimensions. Some of the team related metrics are listed in [64], [65]. Out of these PPDA is perhaps the most informative metric in terms of teams' defensive prowess while most of the other metrics focus on offensive abilities. These metrics are useful in quantifying performance for both teams and players, however, there is more room for describing this complex game and there is need for further rigorous research into the usefulness and development of the metrics.

2.3 Economic and Financial Modelling

Similar to athletic performance, finance in football has multiple layers. The lowest granularity of finances in football is related to player transfer fees and salaries. However, there are also other financial aspects clubs deal with on a regular basis and these are reviewed in a separate section.

2.3.1 Player Valuation

Surprisingly, the most established estimation of player transfer fees is crowd-sourced. This crowd-sourced information is provided by TransferMarkt [30], a large website for Football related information. The crowd-sourcing mechanism works by collect-

ing estimations from individuals, in this case fans, and combining them into a final estimation [66].

Historically, studies on player valuation have been exploratory or descriptive in nature. Initial studies on the topic focused on identifying the factors that affect the market value of the players. These models provide valuable insights for clubs and the policy makers, and they shed light on transfer market dynamics. These insights can further be used to make strategic decisions and make policies in the case of governing bodies.

In [17], a 6-year long study was performed on English Premier League (EPL) matches to understand the effect of player performance, buying and selling club characteristics and time on transfer fees. They use linear regression to perform modelling. They use performance segments as a control variable and find that transfer fees are volatile even within the segments.

The presence of volatility led authors to model players as volatile assets in [18]. They use club's own financial situation and objectives to model the return on investment on an inbound player transfer. This approach values the player specifically for the purposes of the buyer, instead of estimating a generic market value. This approach is far more useful for individual clubs as an investment tool when they have shortlisted a couple of players (i.e. assets) in which they would like to invest. However, this approach cannot scale to the number of available players in the market.

Conversely, predictive models aim to estimate player market values using up-to-date information *prior* to the transfers. Their objective is not to explain the dynamics of the transfer market, but to capitalize on the existing information to estimate player values at any point in time with previously unobserved data. The de-facto benchmark for these estimations is crowd-sourced values by Transfermarkt [66].

In a recent study, authors use an expanded set of variables, including players' names and ids, to estimate transfer fees [19]. While the player's personality is certainly a factor in the case of super-stars in their market valuation, this level of specificity in independent variables makes generalisation of the model impossible, therefore this approach has limited real-world utility. In areas of study where machine learning is an established tool for estimation, researchers are diligent to achieve generalisation. However in a field as new as sports analytics, the rigour has not yet been established and core principles of machine learning are not yet fully settled in practice.

In [67], authors apply linear regression modelling to estimate player transfer fees for top-5 European Football competitions. They use simple player performance statistics and player popularity measurements between 2009 and 2014 to arrive at estimations. This is the first study that includes player popularity in a partially generalisable fashion. They achieve only partial generalisation because they use the final statistics for web-hits from Google, Facebook and Wikipedia for all transfer fees. That means, they use the player popularity measures post 2014 for all of the transfers between 2009 and 2014, which results in an undesirable effect of data leakage [68]. Ideally, for all features and variables included in modelling, the domain of the data must be

equivalent. This concept is the basis for the "Scaling" methodology proposed in this thesis. The proposed approach underperforms compared to the crowd-sourced estimations of TransferMarkt.

In [69] and [70] authors utilise advanced machine learning techniques XGBoost [71] and Lasso regression [72] to estimate transfer fees for a larger set of competitions using data from the video game Football Manager [73]. They use model ensembling techniques to combine predictions from different models and report against TransferMarkt predictions. Their estimations outperform crowd-sourced predictions. However, their approach relies on extensive feature engineering and data transformations already applied by Football Manager developers and is not applicable to real world datasets where majority of the work is related to data cleaning, alignment and feature engineering.

Similarly, in [74] authors use data from the video game FIFA [75]. Instead of estimating individual transfer fees, authors use a clustering approach to estimate a group value for players that fall into the cluster. The estimated market value of the players is the value estimated for the centroid of the cluster. They estimate crowd-sourced market values instead of transfer fees.

2.3.2 Financial Modelling for Clubs

In [7] authors introduce the so-called 'superstar effect' and demonstrate its existence in Italian football. According to this phenomenon, regardless of the size of the difference between the best and the rest in terms of actual performance, the best of the best have disproportionate gains.

In [76], authors expand upon the efficiency of the clubs in their spending and find that all clubs in EPL are inefficient in terms of their investment to various degrees. They demonstrate that some of the most efficient clubs are of smaller size, with a limited budget.

To further analyse the financial decision making of Football clubs, authors employ "Econometric Frontier Modelling" in [16]. They find that clubs in EPL experience diminishing returns on their financial investments.

2.4 Applications of Computer Vision

While not immediately relevant to this thesis, the field of sports analytics stands to benefit immensely from advances in computer vision research. Video analytics is already a staple of the data collection process. In addition to this, as a spectator sport, Football stands to increase viewership experience with intelligent applications of computer vision technology, including but not limited to increasing accessibility for the visually impaired. While video analytics is out of scope of this thesis, comprehensive reviews on the topic can be found in the following papers: [77]–[80].

CHAPTER 3

RESEARCH DESIGN

3.1 Introduction

The main research objective of this thesis is to come up with a framework that combines player performance and player transfers in a data-driven fashion at a global scale. Additional research objectives are to examine data fusion opportunities, develop robust statistical measures, assess the use-case and suitability of different machine learning approaches to such a complex dataset.

In this section, I first describe the research questions posed in Section 1 and then provide the details of the undertaken research.

3.2 Research Motivation and Research Questions

Football analytics so far has been limited to simple statistics and aggregations, however, these approaches are now being replaced with richer and more complex methods.

Established analytics methods in football are limited to performing analyses on a small subset of comparable data. One of the big challenges is to ensure the data from different domains is comparable and, therefore, can be used in analyses together. However, it is known that in football, comparison across different competitions using the same aggregated statistics is not effective, especially if the competitions are different in terms of quality. This results in statistics themselves being incomparable across domains and therefore the analyses are impaired by this limitation. One of the main objectives of this thesis is to identify, then quantify, the factors that must be included to align statistics such that they are comparable across domains.

In addition, statistics in Football are either in the form of frequencies or percentages. However, these two statistics cannot be separated from each other due to a phenomenon called "effect size" [81]. Effect size quantifies the robustness of the statistic by combining the percentage and the frequency together. For instance, 100% shot accuracy in a single shot does not provide any robust information, however, 70% shot accuracy in 10 shots provides more information in the player's ability to score a

goal. These two statistics differ in their effect sizes. Ideally, analyses should reflect this distinction, however, existing methods provide these statistics separately. This thesis aims to propose methods to combine these two in a robust manner.

Furthermore, player performance rankings are dependent on more than the individual performance statistics of any given player. There is also the component of teams' play styles and requirements. Two players playing in the same role can be equally good while having completely different play styles. One of those players could fit the club requirements and the other may not. The ability to modify the search criteria for players is an important contribution. In fact, this preference based query results have been studied extensively in information retrieval domain [82]–[86], however, this has not been applied to sports and other domains. This thesis aims to build a ranking methodology where different requirements can be set and the methodology can be applied to different use-cases.

In Football, there are clubs of all sizes and performance levels. Club financials also vary vastly, therefore different clubs can afford different levels of players. Given the financial constraints and performance objectives of clubs, clubs have an interest in maximizing athletic performance and minimizing the transfer expenditure. In addition, scouting players and contractual negotiations are time-consuming and require the involvement of club personnel. In light of this, the combined picture of player performance and fee is an important tool for guiding efforts. Another application objective of this thesis is to build a comprehensive predictive model to estimate player transfer fees. As a research question, we explore the applicability of various transformations and machine learning problem formulations on transfer fee modelling. The aim is to introduce new approaches to modelling transfer fees and comparing them for further formal studies.

3.3 Player Performance

Quantifying player athletic performance in team sports is a challenging problem. The team success depends on the collective performance of the players, and players' individual performances are dependent on their teammates' performances. The clearest example of this is the performance of the goalkeeper. If the team defence is good enough, or team is disproportionately stronger than the opponent, the opponent may not even have a shot on target which the goalkeeper needs to save. On the other hand, if the defence is not well-formed, then even though the goalkeeper is good, they may eventually experience sufficient number of shots on target that concession is inevitable. Likewise, the performance of forwards is dependent on their midfielders' and wingers' ability to construct chances.

Evidently, each position has different responsibilities. Even within general positions player responsibilities may differ based on sub-positions. To add to this complexity, players may not be able to perform these responsibilities if counteracted by the opposing team. This typically depends on the relative quality of the opponents.

Players' performances in a game also depend on how long they get to play in the match. For example, in certain scenarios, coaches may opt to substitute a specific player in the last few moments of a match, either to strengthen the attack or the defence. In case of strengthening the attack, a last minute goal may be expected of the player despite their low-duration play time. However, if the same player were to play the full match duration, it would be unreasonable to expect a linear increase in the number of goals they score with the play time they are given. Therefore, comparison of performance in different games for a single player is also a nuanced problem.

Typically, player performances are calculated across multiple games in a single competition across a season and to quantify their per 90 minute statistical averages are used [15]. This however, does not solve the problem of comparing players from different competitions. Furthermore, individual players may move to another competition when they are transferred. Simple statistical averages do not translate between competitions and are therefore inadequate to quantify player performance.

Finally, players performances are not constant and may develop or degrade in time. More recent performance of the player is more important than the historically old performances.

In this thesis, we aim to combine not only the duration of game-play, which is already established as an adjustment factor, but also additional contextual factors to extract the raw performance of the player. Therefore, the proposed methodology must be on a player and match level to account for contextual differences in players' matches. For instance, playing against a weaker opponent would result in inflated in-game statistics. Likewise, individual matches in higher-quality competitions are not comparable to lower-quality ones.

This thesis proposes a framework to align in-game statistics to achieve comparability across competitions, play-times and match difficulty. The framework also provides means for time-dependent quantification. In-game statistics develop from the context in which the player plays in, as well as their raw performance. To extract player's raw performance from context-driven statistics, context present in the statistics must be explicitly represented. This thesis first introduces an extendable methodology for context representation, then extends the existing statistical aggregations to include the effect size to propose an adaptable ranking scheme to fit various use cases.

The flow-diagram for ranking methodology is given in Figure . All computations up-to the ranking step are the same. At the ranking step, this thesis employs various techniques to address the ranking scenario. In all cases, individual features on which the player is analysed, so-called "analysis dimensions", are reduced to a single vector of performance. These analysis dimensions depend on the players' roles. For example, a defender would not be analysed on the number of goals they score. The relevant dimensions for various positions are explained in [23].

In the case where club requirements and preferences along those analysis dimensions are known, the weight vector attributed to these preferences is used to reduce the analysis dimension matrix into a vector of performance ratings. Therefore the opera-

tion is a weighted average procedure. In the case where we have an objective function based on club's targets such as maximizing revenue, the weight vector for the analysis dimension is discovered through non-linear optimisation, and then the same weighted average procedure is applied. Finally, where we have no information, we make use of the information retrieval literature and use unsupervised ranking methods [86].

The notation used for player performance ranking methodology is given in Table 3.

3.3.1 Context Representation

To compare the in-game statistics collected from different competitions for different players under varying play conditions, these statistics must first be aligned on the same scale of comparison. This alignment is performed through representing the context from which the statistic arises in mathematical terms. This thesis calls this mathematical representation a "scale" and the process of aligning the statistics is called "scaling".

Given measurements m and adjustment scale s , alignment is defined as in Eq. (1), arriving at a scaled measure m' , through element-wise multiplication.

$$m' = s \odot m \quad (1)$$

In this thesis, I introduce four scales to align player statistics. The number and the domain of scales can be extended to include other factors or the approach can be adjusted to other sports where context is different. These four scales correspond to player's play time in a given match, the difficulty of the match in terms of relative opponent strength, the athletic quality of the competition player is playing in and the recency of the game.

3.3.1.1 Scaling for Play Time

Play duration of the player affects the statistics they obtain in a game. While the frequency related statistics monotonically increase as the player plays longer, the ratio related statistics converge to a match average for the player. To arrive at a performance devoid of the dependency on the time, player's play time must be represented mathematically. However, the time dependency is not linear.

Figures 1 and 2 respectively show the duel frequency and duel success rates for Emre Akbaba (Galatasaray Sports Club) in his match against Yeni Malatyaspor on Jan 12, 2021.

Time dependency in both event frequency and success rate is clearly observed in these figures. Monotonic increase of frequency of actions across time is trivial, however convergence of ratios may not be as easy to imagine. Figure 2 shows the convergence

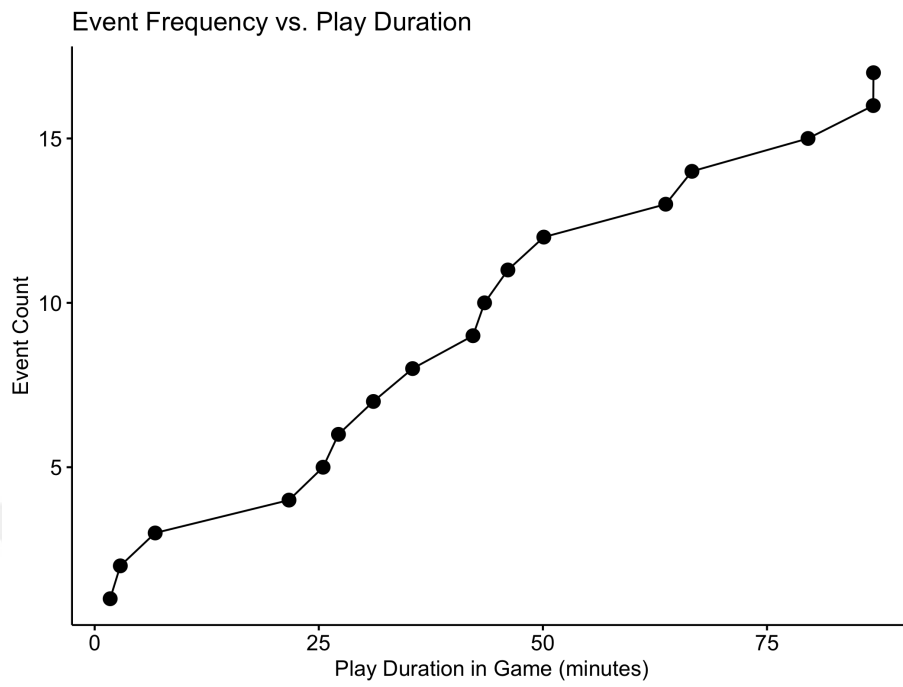


Figure 1: Frequency of Duels

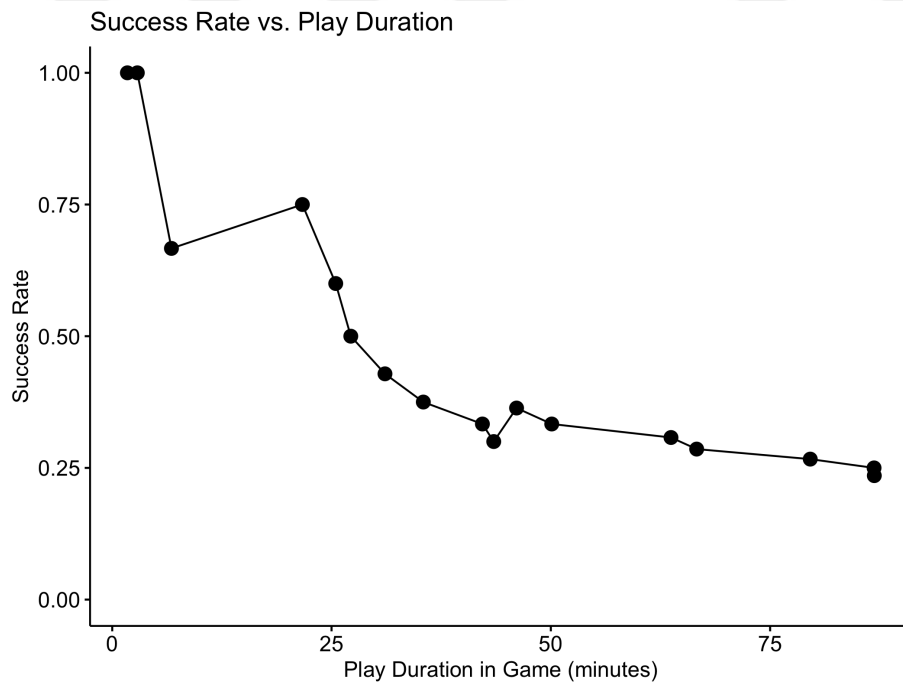


Figure 2: Success Rate of Duels

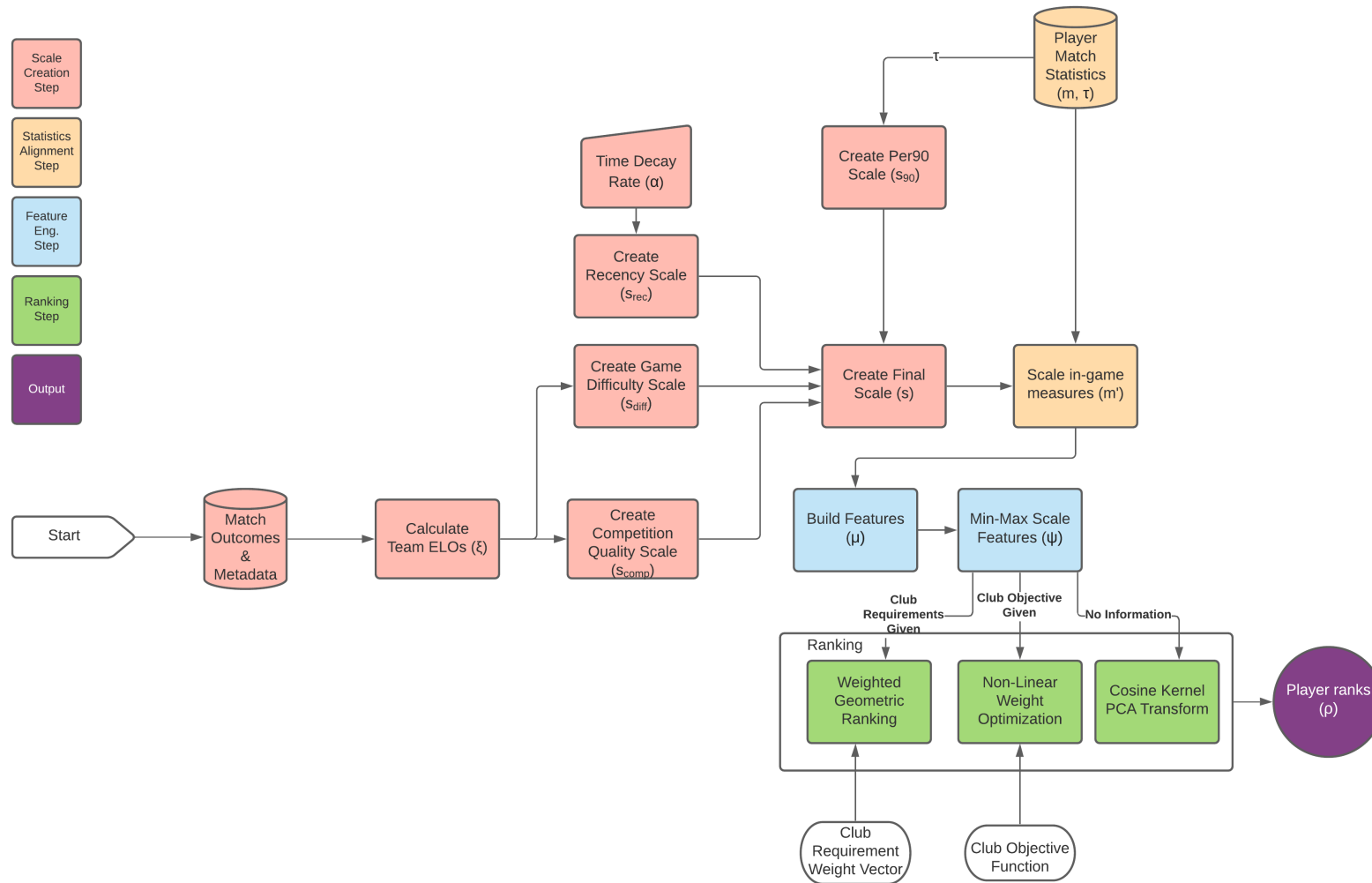


Figure 3: Flow Diagram of Ranking Methodology

Table 3: Definitions and Variables

Term	Definition	Notation	Dimensions
Match	Event of two games competing against each other, from which player performances are extracted	N	N
Player	Players to be rated	p	p
Team	A football team that is active during the analysis period	λ	λ
Analysis dimension	The player measurements that are relevant to analysis. These are the variables that the user wants to measure the performance on.	d	User defined
Measure	In-game statistics collected by the data providers such as WyScout and OptaStats	m	$p \times N \times d$
Minutes played	The duration of player appearance in a match	τ	$p \times N$
Time Decay Rate	Value that specifies the importance given to older matches	α	Scalar
Team Ratings	ELO ratings of teams	ξ	$\lambda \times N$
Scale	External factors that are relevant to player performance which are not directly visible in measures	s	$p \times N$
Scaling	Process to transform the measures to incorporate the different factors that affect the performance. This is a scalar multiplication of the scale and the measure	fs	
Scaled Measure	In game statistics, adjusted for factors through scales	m'	$p \times N \times d$
Aggregated feature	Features computed from scaled in-game statistics m'	μ	$p \times d$
Dimension rating	Player's performance rating on each individual aggregated feature μ	ψ	$p \times d$
Time	Day/Month/Year	t	t

of duel success of player to 25% and stagnating around this value for the remainder of the game.

In the literature, to account for this non-linear behaviour, traditionally, the in-game player statistics m for a single player in a single game are transformed into per 90-minute statistics using Eq. (2) [15].

$$m' = \frac{\sum_{i=1}^{N_p} m_i}{\sum_{i=1}^{N_p} \tau_i} 90 \quad (2)$$

Where τ_i is the minutes played by the player in game i and N_p is the total number of matches played by player p , m is the in-game statistics of player p and m' is the scaled measures of player p .

Visually, the increase in duel frequency and decrease in success rate over time appear to have logarithmic properties. Figures 4 and 5 show the relationship between log transformed play duration of the player and the statistics. Log-transformed play duration has Pearson correlation of $0.88(p = 2.5e - 6)$ and $-0.95(p = 3.6e - 9)$ respectively for frequency and success rate. This correlation coefficient quantifies the linear relationship between two variables, in this case both coefficients show almost perfect linear correlation with the log transformed duration. For the frequency, log transformed duration has a positive correlation with the statistic and for the success rate, there is a negative correlation with an implication that the frequency increases with time and the success rate decreases.

Whether the success rate increases or decreases with time (or remains static) is dependent on the player, however linear relationship with log transformation incorporates the expected impact of play duration and therefore guides the scale proposed in this thesis.

The requirement to develop scales on a per player and per match basis led to the development of a logarithmic per 90 scaling given in Eq. (3).

$$s_{90} = \ln\left(\frac{90}{\tau_i}\right) + 1 \quad (3)$$

Term $\frac{90}{\tau_i}$ quantifies the percentage of play time of the player for match i . Even though Football matches have a duration of 90 minutes, often games go longer due to stoppages or they may even go to overtime. Therefore, frequently players play more than 90 minutes, in which case this term would be less than 1. This would cause the logarithm to be negative. To avoid this issue +1 term is added to the equation. In case of exactly 90 minutes play time, the scale would be 1.

This approximation is empirically shown to have almost perfect linear correlation with outputs of Eq. (2) in Figure 6. This logarithmic formulation is closely related to

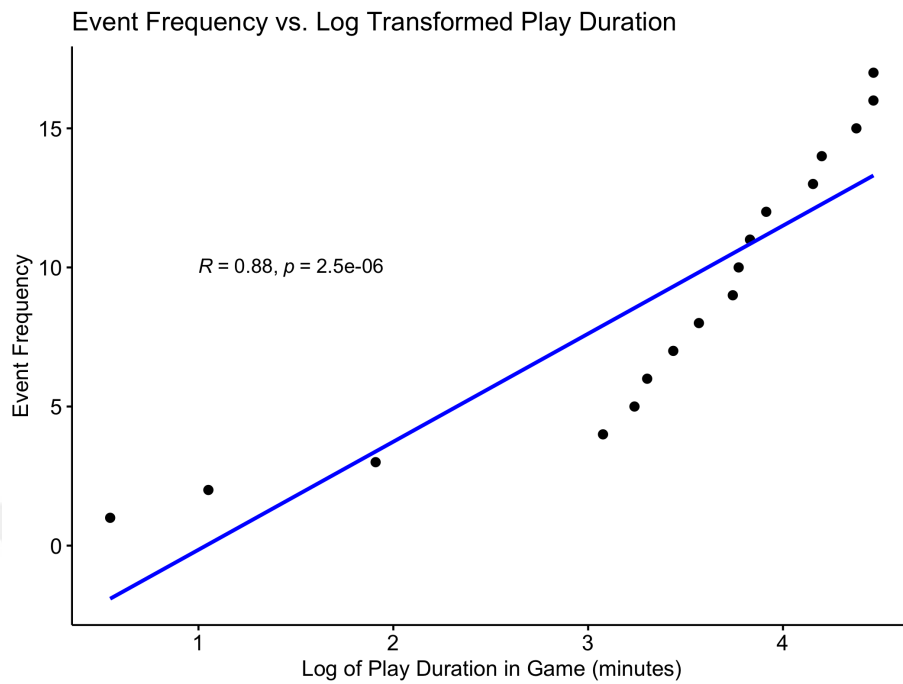


Figure 4: Frequency of Duels over Log-Transformed Play Duration

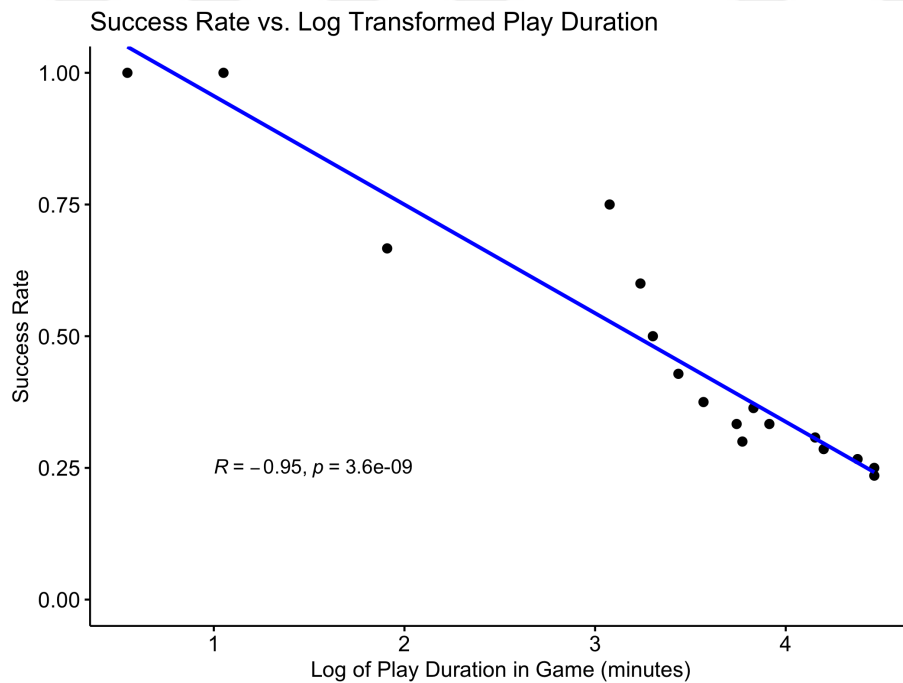


Figure 5: Success Rate of Duels over Log-Transformed Play Duration

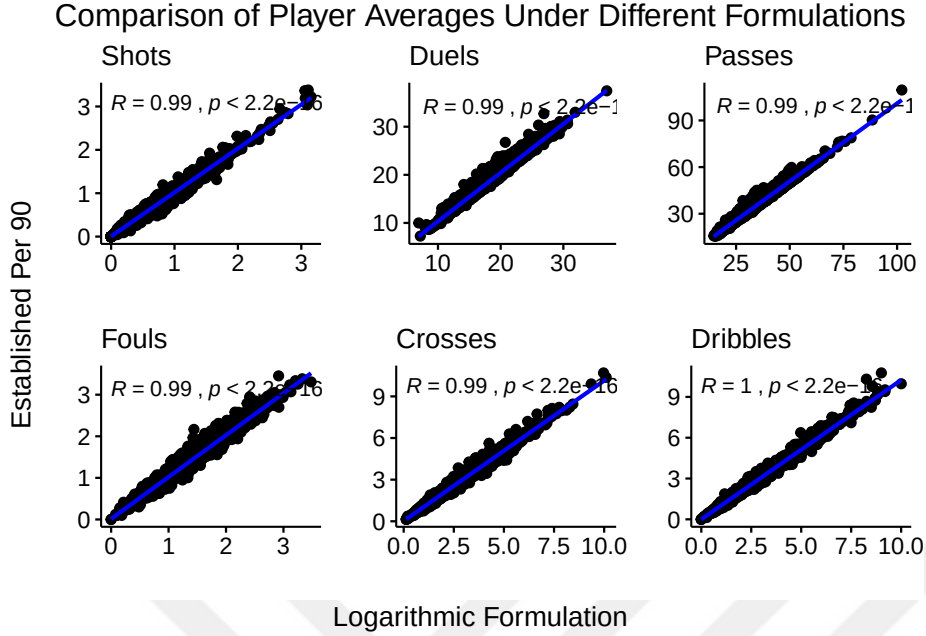


Figure 6: Proposed Scale vs. Established Per 90 Estimations

the generalized harmonic mean approximation provided by [87]. The approximation we provide and the approximation provided in [87] are both based on the relationship between the harmonic mean and logarithms. However, their derivation is theoretical whereas we arrived at this representation empirically. For other types of games with fixed durations, such as Basketball, maximum duration of 90 would be replaced by the maximum duration of the corresponding game.

3.3.1.2 Scaling for Game Difficulty

Player performance in a game also depends on the strength of the opponent. Players who consistently perform well against tougher opponents as well as weaker ones are likely better players. For a consistent and comparable measure of team strengths, this thesis makes use of the existing methods devised for chess [53], [88]. These ratings have also been used in football in isolated competitions [52], [54], [56]. However, existing methods are designed to provide team ratings in isolated competitions, therefore do not generalise to all teams except for minimal cross-competition exposure [54]. To address this limitation, this thesis adopts a data-driven initialization scheme. In [52], authors show that ELO ratings are correlated with team value and as shown in Figure 7, we find that average player values are log-linearly correlated to the UEFA coefficients. This thesis uses these log-transformed average player values to initialize the team ratings, and use the score of each match for each team to update the ratings for the teams via the ELO algorithm [53].

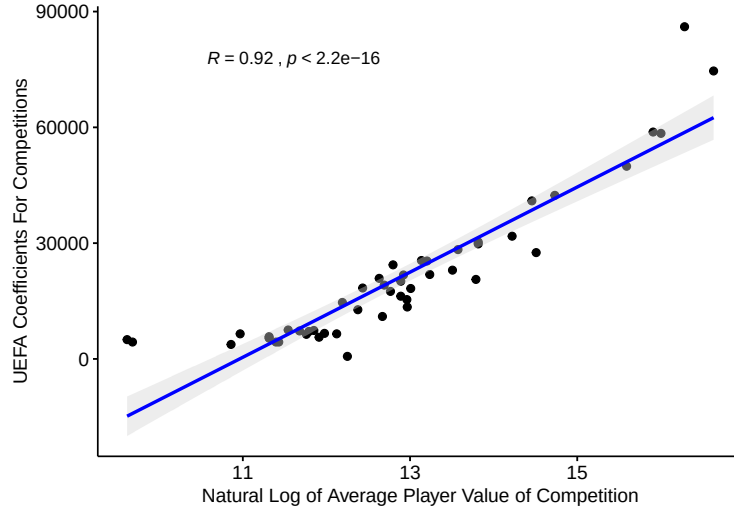


Figure 7: UEFA Coefficients vs. Player Market Values

ELO algorithm takes into account the relative ratings of two head-to-head teams and increases the rating of the winning team based on update parameters and decreases the rating of the losing team. The algorithm uses one parameter, K , which determines the expected level of uncertainty in the estimates. The calibration between competitions is further reinforced through the inclusion of international games. Given the team ratings, game difficulty scale for team λ in a single match is computed as in Eq. (4).

$$s_{diff_\lambda} = \frac{\xi_{-\lambda}}{\xi_\lambda} \quad (4)$$

where, $\xi_{-\lambda}$ is the opponent rating for the most recent match (as the ratings change after every match) and ξ_λ is the rating for the team λ . The parameter values for ELO algorithm are provided in Section 4.

3.3.1.3 Scaling for Competition Quality

There is a difference in the quality of the football played across different competitions, which are reflected in the UEFA Country and club ratings [31]. This makes in-game statistics obtained in different competitions incomparable. However, there is no publicly available global rating of competitions with different divisions, and as a proxy, we first compute the ELO ratings of teams [53] and represent the competition quality as the average ELO of the teams in a competition. As opposed to the game difficulty scaling where the statistics are adjusted to account for the strength of the opponent, competition scaling boosts the statistics obtained in higher rated competitions to reflect the overall quality of game play. Eq. (5) shows the competition scaling

as the average of ELO points of all teams in a given competition where Λ_{comp} is the number of teams in the competition.

$$s_{comp} = \frac{\sum_{\lambda=1}^{\Lambda_{comp}} \xi_{\lambda}}{\Lambda_{comp}} \quad (5)$$

3.3.1.4 Recency Scaling

When evaluating performance, the recent games are more important than the older ones. However, effect size is important in statistical analyses and a sufficient sample size is required for statistical inference. In this study, we adopt a decaying mechanism limiting the impact of older games on the statistical results. Exponential decay mechanism in [34] is based on incremental averages that result in an additional subtraction of higher order decays. In contrast, our approach aims to capture variances in each individual game (like in per 90 scaling), so we rely on being able to scale each game individually as opposed to incremental averages. Recency of the game is encoded using Eq. (6).

$$s_{rec} = \alpha^{\Delta t} \quad (6)$$

where α is a real-valued number $0 < \alpha \leq 1$ and Δt is the time difference in weeks from the latest game to the date of analysis.

3.3.1.5 Combining The Scales

In-game statistics for all players need to be scaled according to different factors that affect the game. Computation of the final scale is performed by multiplying all the computed scales for the player. For a player p , the final output for an in-game statistic (i.e. measurement) m , would be computed as in Eq. (7).

$$m' = s_{90}s_{comp}s_{diff}s_{rec}m \quad (7)$$

All scales are player and match specific, based on the player's minutes played in a given match, the competition the match is played in, the difficulty of the match, as well as the recency of the match as defined in relevant sections. The scalar production of the scales allows for toggling them on and off as relevant and needed. For instance, if the analysis is restricted to only one competition, there is no need for including competition scaling. Or, in case we wanted to compute "the best player of all times", the recency scale could be completely toggled off.

3.3.2 Statistical Aggregation

To arrive at player-level statistics, in-game statistics for each player must be aggregated into a single value. Aggregation is applied after the scaling step and we employ three basic techniques: scaled averages μ_{avg} (Eq. 8), scaled ratios μ_{rat} (Eq. 9) and strength variable μ_{str} (Eq. 10). Given the definitions of an analysis dimension d , these aggregated features are computed as in respective equations 8, 9, 10.

Eq. (8) reflects the scaled averages of in-game frequencies such as fouls performed or shots taken. This quantifies the activity level of the player in a game. Eq. (9) reflects the scaled performance measures such as recoveries compared to fouls in the form of ratios. This variable quantifies the relative activities between two variables, or when the numerator is the successful subset of the denominator, success rate. Finally, Eq. (10) combined the ratios with the sample size to arrive at a better rounded success measure.

$$\mu_{avg_{pd}} = \frac{\sum_{i=1}^{N_p} m'_{1_{ip}}}{N_p} \quad (8)$$

$$\mu_{rat_{pd}} = \frac{\sum_{i=1}^{N_p} \frac{m'_{1_{ip}}}{m_{2_{ip}}}}{N_p} \quad (9)$$

$$\mu_{str_{pd}} = \frac{\sum_{i=1}^{N_p} F_{binom}(m_{1_{ip}}, m_{2_{ip}}, \bar{\pi})}{N_p} \quad (10)$$

In these equations, $m_{j_{ip}}$ is the j^{th} measure for player p , F_{binom} is the Binomial cumulative distribution function, $\bar{\pi}$ is the average ratio across all players of the two measures in Eq. (9). When calculating μ_{rat} , the measures do not necessarily need to be related. For instance, $m_{1_{ip}}$ and $m_{2_{ip}}$ may refer to fouls and recoveries respectively. However when calculated μ_{str} , $m_{1_{ip}}$ must be the successful subset of $m_{2_{ip}}$, because this aggregated feature aims to quantify the amount of evidence towards success in an analysis dimension. Therefore, the strength variable is an extension and improvement over success related ratios in the sense that not only the success of the action is quantified but also the *effect size* [81]. For instance, a ratio of 80% with 5 samples would have less strength than the same ratio of 100 samples, which provides more evidence towards the expected value of the ratio.

There are known issues regarding averaging ratios in μ_{rat} , pointed by [89]. The scaling approach mitigates these issues by transforming them into a non-ratio range.

Finally, computed aggregated statistics must be standardised before further use, as they have different ranges. Therefore, they are standardised using min-max standardization F in Eq. 11. This standardization is called *dimension rating* ψ_d for analysis

dimension d , which quantifies the performance of the player in this dimension. The result is always in the range of $[0, 1]$.

$$\psi_d = F(\mu_d) \quad (11)$$

3.3.3 Ranking

Final step in the proposed methodology is ranking. In the context of this thesis, ranking step can take three different forms. These forms depend on the level of preference information specified by the ranker which is usually a club.

Clubs employ technical staff as well as analysts who may be able to distil the requirements for the player searched into numerical terms. For instance, when looking for a forward, they would be able to provide relative preference information such as "shots on target ratio is two times more important than success ratio in aerial duels". Given the relative importance weights in all dimensions relevant to the search of a player, the ranking step would simply become a vector/matrix multiplication operation, where the vector is the relative importance weights provided by the clubs.

However, articulation of such preferences in terms of numbers is not straightforward. As an alternative, the ranker could provide a general objective such as increasing their sportive success (i.e. ELO rating) or their revenues and then, an objective function could be used to find the set of optimal weights for the final ranking step, using non-linear optimisation techniques. While the rest of the computation steps are still the same, in this case, the weights are driven by a function eliminating the need for ranker-provided weights.

In addition, rankers having an exact preference is a rare case and, they may not have an objective in mind either, where a generally applicable output is desired. In such cases, the ranking methodology employs unsupervised learning techniques to arrive at the final ranks.

3.3.3.1 Preference Based Ranking

In the next step, dimension ratings must be combined to a single metric that quantifies the overall player performance. This step is performed by taking the weighted geometric average (Eq. 12) of all dimension ratings. Geometric average is chosen as the aggregation function because arithmetic averages are sensitive to outliers. The weights are user defined parameters that represent the importance of each dimension in the analysis.

$$\rho_p = \exp\left(\frac{\sum_{d=1}^D w_d \ln(\psi_d)}{\sum_{d=1}^D w_d}\right) \quad (12)$$

In this equation ρ_p is the final rating for player p , w_d is the expert specified weight for analysis dimension d .

3.3.3.2 Rankings Towards an Objective

In cases where preferences are not explicitly provided, the weights can be deduced from the data. This process would require the definition of the objective function, f_{obj} , and an optimisation process that uses part of the data to iteratively estimate the current value of f_{opt} and optimises the weights by measuring the change in performance. This process is defined in Algorithm 1.

Algorithm 1: Weight Optimisation Procedure

Result: Set of optimal weight vectors W^*
Data: tol_{max} , $iter_{max}$, $numBins$, P , f_{obj} , μ (Table 3), d (Table 3), $trainSetSize$, $validSetSize$, $testSetSize$
 Create the training set P_{train} ($trainSetSize\%$), the validation set P_{valid} ($validSetSize\%$), the test set P_{test} ($testSetSize\%$) split from set of players P where players have post-analysis transfers, without replacement;
 Calculate f_{obj} for each player p in P ;
 $iteration = 1$
for $iteration \leq iter_{max}$ **do**
 Initialize weight vector W of length d , where each $w \in W \sim \mathcal{U}(0, 10)$
 Initialize $tol = 1e10$
 while $tol > tol_{max}$ **do**
 $R = W\mu^T$ Calculate player ratings;
 Discretize R into $numBins$ rank groups RG of R using equal frequency discretization ;
 Subset $[RG_{train}, RG_{valid}, RG_{test}]$ for $[P_{train}, P_{valid}, P_{test}]$;
 Calculate average value change in objective function $AVC_{train}[i]$ for each rank group i in RG_{train} ;
 $AVC_{train}[i] = \frac{1}{N_i} \sum_{j=1}^{N_i} f_{obj}$ where N_i is the number of players in rank group i in RG_{train} ;
 $RC_{train} = Cor_s(RG_{train}, AVC_{train})$;
 $W, tol = f_{opt}(RC_{train})$ where f_{opt} is L-BFGS-B Algorithm ;
 end
 $AVC_{valid}[i] = \frac{1}{N_i} \sum_{j=1}^{N_i} f_{obj}$ where N_i is the number of players in rank group i in RG_{valid} ;
 $RC_{valid}[iteration] = Cor_s(RG_{valid}, AVC_{valid})$;
 $iteration = iteration + 1$;
end
 $W^* = argmax_w(RC_{valid})$

3.3.3.3 Rankings with No Feedback

In cases where there are neither preferences, nor a mathematically-defined objective function, the rankings are computed using unsupervised learning methods. This thesis performs this step by using a dimension reduction approach, Kernel PCA with cosine similarity. In [86], authors specify requirements for a suitable ranking approach and offer PCA as a simple linear solution. However, they also mention that PCA requires orthogonality of the components and therefore it is not suitable for most real-life ranking problems. Kernel PCA is presented as a solution to non-orthogonal multivariate ranking problems. They take the solution a step further and use Bézier curves for one-dimensional embedding. However, Bézier curves require a guarantee for no-ties and in our case we cannot guarantee such topology. Therefore, this thesis employs Cosine Kernel-PCA, which is an established method in document ranking tasks.

3.4 Player Valuation

In contrast to exploratory and descriptive methods that lack generalisation, predictive modelling approaches aim to estimate player values in the market given the available information at the time of estimation. This constraint on the timeliness of the information limits predictive modelling approaches to use only information available prior to transfers. The generalisation requirement also prevents predictive modelling approaches from using any unique identifiers such as player ids and names. The idea of generalisation is being able to use the same already-trained model on new data which is unobserved prior to estimation and still being able to have reasonable estimations. Assume we have a "predictive" model which uses player names, and a new young player bursts into the scene, existing model would not be able to estimate the market value of this player until they have had at least one transfer. In a predictive setting, this would defeat the point whereas it is completely feasible in an exploratory setting in which we aim to understand the factors that affect valuation.

Naturally following from this restriction, predictive models omit variables and information post-transfer and unique ids. Therefore in a predictive setting with the exact experimental setup, we expect descriptive models to have better fit to the data. However, with the advances in machine learning, predictive models now achieve good performance on estimations making them feasible to use for real-life applications.

The objective of this thesis is to build a comprehensive framework that has real-life applications. Therefore, instead of elaborating on already rich exploratory and descriptive literature on player values, we opted to build a predictive solution using a large, real-world dataset.

The research design is guided by the research questions outlined in Section 1 and literature review of transfer market dynamics. To answer these questions, I first collect data from various sources. The raw in-game statistics are provided by WyScout [29], transfer and financial information are scraped from Transfermarkt [30] as well as in-

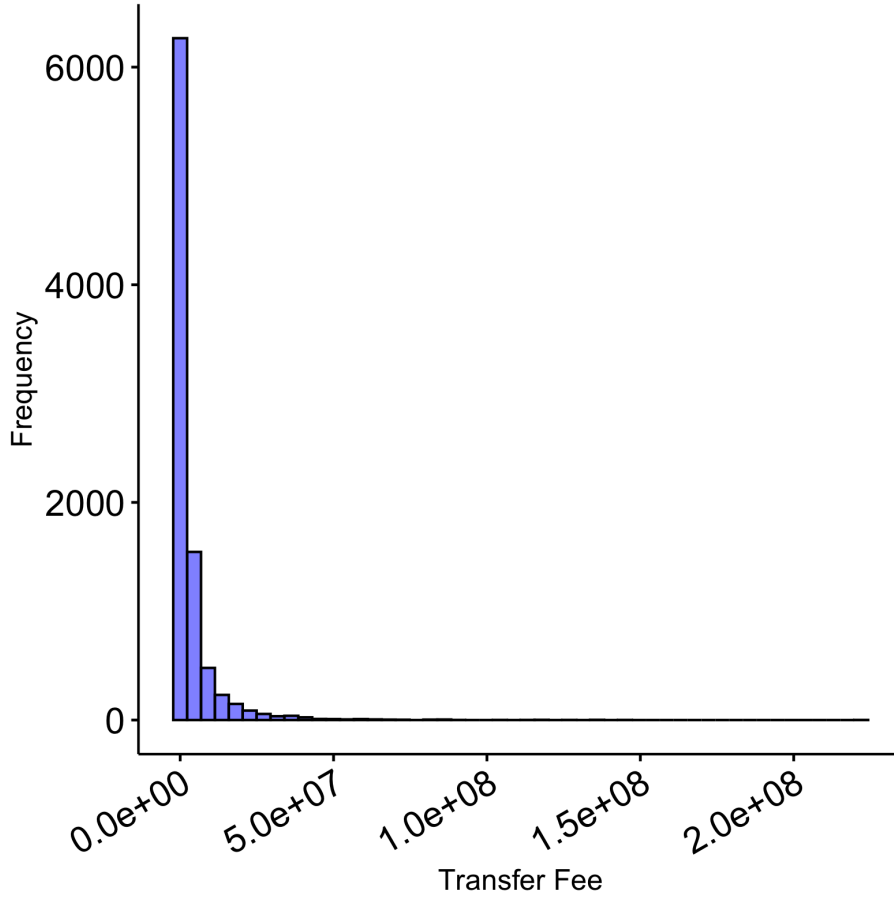


Figure 8: Distribution of Transfer Fees

juries. Finally, we collect Google Search [90] trends to include the historical player popularity information.

3.4.1 Player Value and Market Dynamics

As shown in several studies [7], [11], [16], [76], transfer fees and club spendings are skewed and inefficient. In one study that approaches the transfer market dynamics from a physics perspective, the transfer fees and player salaries are shown to follow power-law dynamics [11]. This means that the transfer fees have a highly skewed distribution. The distribution of all transfer fees since 2015 is given in Figure 8.

3.4.2 Treatment of Transfer Fees

There are several ways of dealing with skewed outcome variables. The choice of the method depends on the use-case of the final solution and the rate of skewness in the data.

In [69] and [70] authors log transform the transfer fees to deal with skewness. In this thesis, we explore several options and outline use-cases where different transformations may be suitable. These are further demonstrated in Sections 4 and 5.

In addition to individual transformations, we also explore the joint-use of continuous and discrete transformations to estimate transfer fees.

3.4.2.1 Continuous Transformations

There are multiple non-linear transformations to further normalise the variable distribution that result in a continuous output. Most well-known transformations in this category are: root transformation, log transformation and Box-Cox transformation.

Out of these, root transformation modifies the distribution the least amount while Box-Cox transformation is the most drastic one. In literature, log-transformation is likely the most common transformation applied to skewed distributions.

This thesis applies log and root transformation to the transfer fees to achieve a distribution closer to Normal distribution. Unlike the other two, Box-Cox transformation requires estimating a model fit to the data and is much harder to extend to the new data. Therefore, we do not consider Box-Cox transformation for the scope of this thesis. Figure 9 shows the effect of root and log transformations on the transfer fees.

Typically, Normal distribution is desired from a mathematical standpoint to model continuous variables as several methods rely on the assumption of Normality. However, methods that do not rely on an assumption of Normality can benefit from noise in the data [91]. Similarly, fitting to a skewed outcome variable presents a trade-off between estimating outliers and estimating general behaviour. In other words, fitting models to outcome variables that are closer to Normal distribution models the general behaviour better, while fitting models to the skewed outcome variables without transformation or with minimal transformation shifts the model towards estimating outliers in the tail better.

In case of transfer fees, there are several high-profile, high-value transfers. Accurate prediction of these transfers is of interest to this thesis because they have high impact in club financials as well as accurate prediction of general cases.

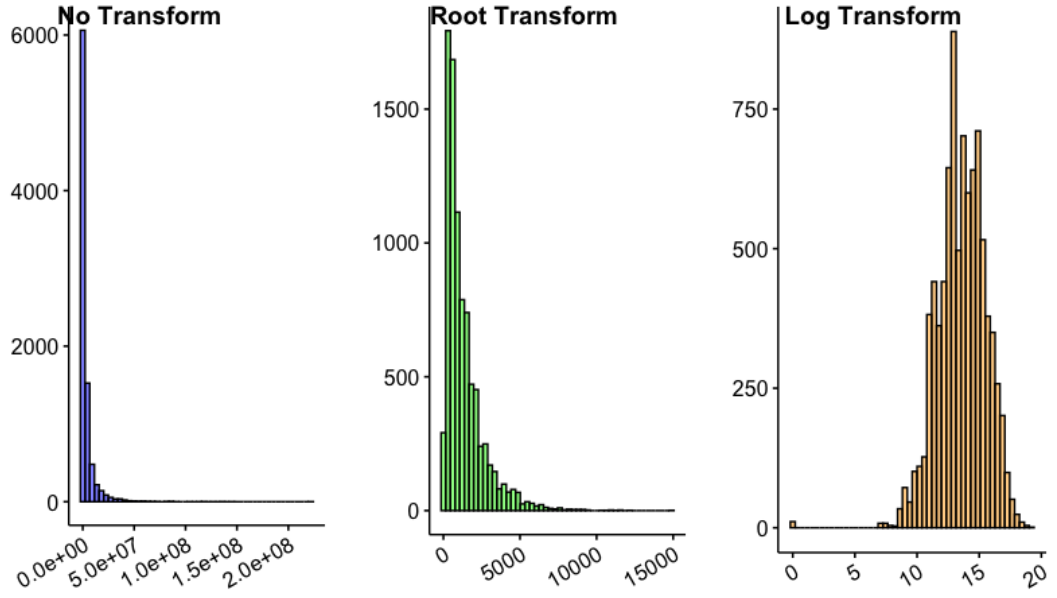


Figure 9: Effect of Root and Log Transformations

3.4.2.2 Discrete Transformations

Another way to deal with a skewed continuous outcome variable is through discretizing the values. By binning the outcome variable to N distinct groups of values, the skewness can be mitigated.

Most frequently used discretization technique is binning the continuous variable into equal-frequency groups. When $N = 4$, equal-frequency binning is called "Quantization" or "binning into quantiles". When $N = 10$, the groups are called "deciles" and when $N = 10$ they are called "quintiles". Equal-width binning is not suitable for such skewed data as the class frequencies for outlier values would be too few.

In addition to equal-frequency binning, "natural breaks" algorithm was proposed in [92]. This method discretizes the data by minimizing the variance in groups and maximizing the variance between the groups. One-dimensional K-means clustering is equivalent to natural breaks algorithm in case of convergence.

When the target variable is transformed into a discrete set of values, transfer fee groups are obtained. Assuming there are N discrete classes from 0 to $N-1$, 0^{th} group would ideally contain the players with lowest estimated transfer fee and $N - 1^{th}$ group would have the players with the highest estimated transfer fees.

Table 4: Major Factors Impacting Player Transfer Fees

Feature Family	Definition
Team Historical Transfers	Transfer behaviour of player’s club prior to transfer. These variables quantify whether the club is likely to sell their players or not.
Player Historical Transfers	Statistics regarding player’s previous transfers. Some players are repeatedly loaned to other clubs to accelerate their athletic development before their team either to eventually utilise the player or to sell them.
Player Demographics	Player demographics such as their age and passport area.
Player Performance	Player’s athletic performance quantified using the feature aggregation and ranking methodology proposed in Section 3.3.
Injuries	Statistics of players’ historical injuries and whether they are injured at the time of relevant transfer season
Player Popularity	Player popularity, quantified using the approach outlined in Section 3.4.3.4.
Team Quality	Athletic performance of player’s club prior to transfer.
Competition Quality	Quality and financial situation of player’s competition prior to transfer.
Team Appearances	Appearances of the player’s club in various international and domestic competitions, prior to transfer.
Player Appearances	Appearances of the player in various international and domestic competitions, prior to transfer.
Player Role	Statistics regarding player’s role in games prior to transfer.

3.4.3 Factors Considered and Feature Engineering

Exploratory studies show that transfer fees depend not only on the player performance but also other factors [17]–[19], [66]. These factors are described in Table 4.

Factors outlined in Table 4 are overarching groups of variables. They branch into several features that represent different analytical variables. These individual features are described as part of an experimental setup in Section 4. In the following subsections, we describe the design of these factors. Player performance quantification has been already described in Section 3.3, therefore it will not be detailed further here.

3.4.3.1 Team Historical Transfers

Transfer behaviour of the current team is also an important factor in transfers. For instance, teams who bought multiple players the previous season may be more inclined to sell players in the current one due to financial restrictions. Similarly, teams who have loaned players in the previous season may want to purchase the players they bought or may want to replace them. Furthermore, teams who consistently buy or sell players may have the financial liquidity to continue the transactions. To avoid

data leakage, the transfer statistics are only calculated for player's club right before the transfer.

3.4.3.2 Player Historical Transfers

Similarly, players' historical transfers are informative of the type of asset they are deemed to be. For instance, a given player could be a staple of the team rosters that they have played in, therefore, they are likely to have fewer loans out, longer duration in their teams and higher-than-average paid transfers. Conversely, an asset under development (i.e. young player) may be loaned out to several clubs before the club holding the contract of the player decides to put them on the field. Furthermore, historical average fee paid to obtain the asset is an informative variable that gives insight to market value of the player. The features are calculated up-to-date of transfer season for any season to avoid data leakage.

3.4.3.3 Player Injuries

Data provider from which the player historical injuries are collected provides 343 unique injuries for all active players. These injuries cannot be used directly. Instead, we group injuries based on how frequently they occur in players, and how severe the impact is. The frequency of an injury is defined as number of unique players experiencing the injury. The severity of the injury is defined as median duration the injury causes the player to miss matches. We group these statistics into decentiles. First decentile being the least common/severe and the 10th decentile being the most common/severe injury.

3.4.3.4 Player Popularity

Player popularity is a major factor in determining transfer value in modern football. To include popularity measures, we collected 5-year long web search data for main football roles (Goalkeeper, Striker, Defender and Midfielder), and the players in comparison. The search trend for their role, search trend for the specific player and the ratio between the two time-series are used as features in modelling. The ratio between the player popularity and the role popularity quantifies the relative popularity of the player compared to the popularity of the role. For example, 'Cristiano Ronaldo' has significantly higher search frequency than his general role 'Midfielder'. Whereas in most cases, the role will have higher search frequency. The main objective of this factor is to quantify the super-star effect.

3.4.3.5 Team Quality

Team quality is defined as club's athletic performance. Their transfer behaviour is already covered in Section 3.4.3.1. Other aspect of the team quality is their athletic performance. This aspect is represented using teams' ELO [53] ratings at the beginning of the transfer season for each transfer season. As ELO is an incremental algorithm that combines historical information as well as the latest performance, there is no need to average team's ELOs within the season. Using the latest rating at the end of the season is enough to reflect the general behaviour in the season.

3.4.3.6 Competition Quality

Much like team quality, competition quality can be defined using ELO ratings. Following from the novel initialization approach to calculating globally comparable team ratings provided in Section 3.3, competition athletic quality can be represented as the average ELO of the clubs in the competition. Furthermore, competitions' financial quality is represented by the average value of the players in the competition.

3.4.3.7 Team and Player Appearances

International and domestic competitions are venues for players to showcase themselves. While domestic competitions allow players to demonstrate their skills to local audience and clubs, international competitions allow for a more global audience. This exposure impacts players' eventual transfer fees.

Similarly, however not as directly, players' clubs' appearances in international competitions and domestic cups also impact players' transfer fees. In case where the player is part of a team with consistent international competition appearances, player is expected to be valued higher than similar players with the same background. Even if they may lack appearances, simply being part of the team provides exposure.

3.4.3.8 Player Role

Player role affects player transfer fees from multiple angles. It has been shown that strikers tend to have higher transfer fees than other players as well as higher salaries [93]. In addition, player's versatility is also another important factor which allows more flexibility in formations.

To represent these factors, we quantify the total play-time percentage in each of the sub-roles. These sub-roles are stated in Figure 10. The definitions of the roles are given in Appendix A.

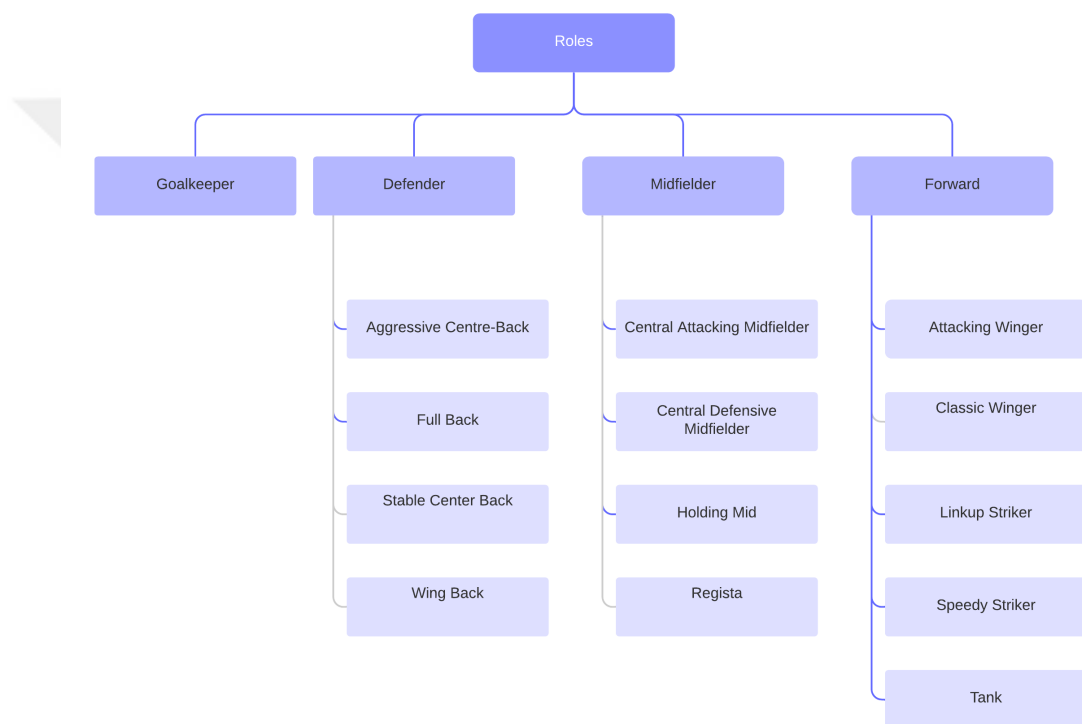


Figure 10: Roles in Football

3.4.4 Modelling

Following from the research questions, we aim to perform accurate estimation of transfer fees using real-world datasets. An additional objective of this thesis is to compare and contrast various modelling approaches. To achieve these goals, this thesis performs the following experiments to decide on the best way to model skewed transfer fees.

This section gives the methodology details for the modelling approach. The exact choice of the modelling algorithm and experimental parameters are provided in Section 4.

Multi-class classification, ordinal regression and multi-label classification all use the discretized values of the transfer fees. However the model training is different. Ordinal regression is trained using a regression over the discretized target. Multi-class classification is the traditional classification approach for modelling discrete target values. Finally, multi-label classification is a different approach to ordinal classification or regression problems.

In multi-label classification, the outcome variable falls into multiple classes at once. Most often, this problem appears in image classification problems. An image taken randomly in nature could be both of an animal and of landscape. In this case, the image would fall into both classes.

In case of transfer fee modelling, the problem of estimating the transfer fee, *group* of the player is translated into "Player achieving at least sufficient amount of transfer fee to fall into group X ". In probabilistic terms, this is a survival function representation such that the aim is to estimate $P(\text{PlayerFee} \geq \text{TransferFeeGroup})$. This means, if a player is valuable enough to fall into the highest transfer fee group, s/he would *also* achieve enough market value to fall into all other groups. To illustrate the transformation, assume we have 5 classes in a problem, from 0 to 4. Then the classes would be transformed as in Figure 11. This method was originally applied to Neural Networks to generalise them for ordinal estimation problems [94]. In a multi-label classification problem, each observation is scored based on their likelihood to fall into each class simultaneously.

3.4.4.1 Model Training and Evaluation

The best practice in building machine learning models is first splitting the data into training and hold-out sets, then using an appropriate cross-validation scheme on the training set to finalize model hyper-parameters. Then the final model generalisation is tested on the hold-out set. In an ideal case, training metrics, cross-validation metrics and hold-out metrics should have comparable values.

Most common method for cross-validation is K-Fold cross-validation [95]. In K-Fold cross-validation, the dataset is divided into K groups, the model is trained on $K - 1$ folds and the performance is evaluated on the remaining fold. The set of hyper-

		Multilabel Class			
		1	2	3	4
Original Class	0	0	0	0	0
	1	1	0	0	0
	2	1	1	0	0
	3	1	1	1	0
	4	1	1	1	1

Figure 11: Multilabel Encoding

parameters whose resulting model has the best cross-validation performance is chosen as the training parameters. These hyper-parameters are used to train a bigger model on the training set, and then finally, to evaluate generalisation, this model is used to predict values of the hold-out set. If the performance on hold-out set is comparable to training and validation scores, and all these scores are above a desired performance level, then model is said to achieve generalisation.

To sum-up the requirements for the simplest form of model evaluation and selection method, we need:

- An evaluation metric and a target for achieving acceptable performance
- A suitable sampling scheme to split the dataset into training and hold-out sets
- A pre-defined K for cross-validation
- A machine learning algorithm that will be used for modelling and its corresponding hyper-parameters
- An initial set of hyper-parameter values for the model hyper-parameters we want to optimise

Perhaps deciding on the target metric and an initial sampling approach for partitioning the hold-out set are the two most important steps in model evaluation.

To illustrate the importance of metric selection, assume we have a very skewed dataset with 100 observations. If 99 of these observations have the class label of 0 and only one is from class 1, and if we simply label everything as 0, we would achieve 99% prediction accuracy. However, that means very little because we were probably trying to detect the cases where class 1 is present. In this case, *accuracy* would be an inappropriate metric.

Similarly with the sampling in case of skewed datasets, if we split the initial 100 observations into 80% training and 20% hold-out set and we include the one sample we are interested in, in the hold-out set, then our model would never learn the pattern that leads to the positive class. Therefore it is important to maintain the pattern of interest in all stages of modelling.

In case of time dependency and multiple observations from the same group across time, modelling and evaluation get even more complex. Transfer fees are shown to have a time component. Not only they occur seasonally, inflation also plays a part in player transfer fees. In addition to this, players have multiple transfers. To ensure generalisation to new players, all data for some players must be reserved for the hold-out set. Therefore, the initial partitioning and the following cross-validation scheme must cover these requirements.

Time-dependent datasets are split using so-called "out-of-time" sampling [95]. In addition, proper distribution of the target variable across splits and partitioned datasets is ensured by "stratification" of the target variable. Finally, generalisation to groups of data (i.e. players) is ensured via "out-of-sample" partitioning.

In this thesis, we employ both out-of-time and out-of-sample partitioning, as well as stratification, to split the full dataset into training and hold-out sets. The cross-validation scheme used is "Grouped Out-of-Time Cross-Validation" in which "time-groups" are used. Each transfer season is a time-group. The folds are generated such that the validation fold is the next season and the training folds are the seasons leading up to the next transfer season at each split.

3.4.4.2 Hyperparameter Optimisation

Cross-validation is the method to find the most optimal hyperparameter values for machine learning algorithms. As the number of distinct hyperparameters grow, the search space for the optimal values increases combinatorically. In such cases searching the entire parameter space is unfeasible. To address this, researchers use random search, or they manually restrict the search space by providing search grids.

Another approach that has gained popularity is "Bayesian" hyperparameter optimisation [96]. In this approach, posterior probability of maximizing the objective function over the hyperparameter search space is estimated at each optimisation step. This posterior probability distribution is then used to select the next set of candidates for optimal hyperparameters. This process is repeated until optimal metric value is achieved or number of maximum iterations is exhausted. In this thesis, I employ Bayesian hyperparameter optimisation to search the hyperparameter space.

3.4.5 Model Ensembling and Information Extraction

It is common for different models to learn different patterns in the data, especially when the data results from complex dynamics. In this case, it is possible to use multiple models together to arrive at a better estimation. In fact, the most prominent algorithms in the literature such as Random Forests [97] or Gradient Boosting Machines [98] rely on this idea. In context of machine learning, combining results from individual predictors is called "ensembling".

There are three main ensembling methods: Bagging [99], boosting [100] and stacking [101]. Bagging technique averages the predictions of individual predictors where the problem is a regression problem. In case of classification problems, it simply performs majority voting. Random forests "bag" the predictions of individual decision trees they are made of.

Conversely, boosting uses an iterative approach to reducing observation-level error. In each boosting iteration, observations that have high error gain more weight for training and the individual predictors are updated with the new weights. Boosting therefore is an observation-weighting method to reduce error.

While not being particularly new, stacking has lagged behind bagging and boosting in terms of popularity. Unlike bagging and boosting, which use relatively simple arithmetic operations to combine the predictor decision, stacking uses *another*, potentially completely different machine learning model to intelligently learn the best combination of predictor decisions. This machine-learning model is called a "meta-learner". There is no standard "meta-learner", the design of stacking approach depends on the researcher building the final solution.

In all cases of ensembling, the performance of the final predictions benefit from "base-learners" capturing different information from data. The difference between base-learners increases the variance in all base learners and reduces bias. Ensembling of the base-learners reduces the variance of final predictions. With low bias from base-learners and low-variance in final predictions, ensembling models perform better than individual predictions.

In this thesis, I use different models to represent the different types of transfer behaviour. Namely, while being few, super-star transfers are of interest for modelling. Therefore to further emphasize these, I adopt a model stacking approach to accurately predict both "regular" and "super-star" cases.



CHAPTER 4

EXPERIMENTAL SETUP

The proposed methodology is applied to real-world datasets, combined from various sources. This section specifies the experimental setup and evaluation choices for each of the components in the thesis.

4.1 Player Performance Evaluation

This thesis applies the proposed methodology to the problem of ranking left or right backs in seasons 2016/2017 and 2017/2018. We apply the methodology to two seasons instead of one season which is more commonly observed in ranking studies. This choice is deliberate to demonstrate the performance of the proposed methodology on longer duration intervals. The season of 2018/2019 is used to demonstrate the real-world performance of the methodology.

4.1.1 Dataset

The experiments are performed using the dataset provided by Wyscout [29], a sample of which is available online [34]. The methodology is applied globally, with no geographical restrictions. There are restrictions on the minimum game-play duration and the number of games. The full set of restrictions is given in Table 5. These filters are aimed at facilitating a sound statistical analysis, because a minimum number of samples per group is required for solid statistical analysis. The only contextual filter is the target position. This is due to positions in football having distinct responsibilities and comparison across positions in terms of performance is unfeasible.

Under these filters, the resulting dataset has 3657 unique left or right back players and 150810 player-match pairs resulting in 1.9 million in-game measurements for 9 analysis dimensions using 13 variables provided by Wyscout. The analysis dimensions, variables used in the analysis dimensions and statistical aggregations are given in Table 6.

In addition, this thesis uses other publicly available data such as transfer values (for ELO initialization). A summary of data sources known (but not necessarily used in this thesis) is provided in Table 2 for the sake of completeness.

Table 5: Dataset Restrictions

Property	Value
Target Position	Left or Right Back
Minimum Play Duration	20 minutes
Minimum Number of Games	10
Minimum Percentage in Relevant Positions	90%

Table 6: Analysis Dimensions

Analysis Dimension	Aggregation	m_1	m_2	Weight
Interceptions	μ_{avg}	# Interceptions		2
Shot Assists	μ_{avg}	# Shot Assists		3
Loose Balls	μ_{str}	# Successful Loose Ball Duels	# Loose Ball Duels	1
Fouls	μ_{avg}	# Fouls		1
Forward Passes	μ_{str}	# Successful Forward Passes	# Forward Passes	1
Progressive runs	μ_{avg}	# Progressive runs		2
Recoveries	μ_{rat}	# Recoveries	# Fouls	1
Defensive Duels	μ_{str}	# Successful Defensive Duels	# Defensive Duels	3
Crosses	μ_{str}	# Successful Crosses	# Crosses	2

Table 7: Parameter Details

Method Component	Parameter	Details	Exp. Value
Game Difficulty Scale, Competition Quality	K	The factor that determines the amount of impact each game has on the team rating adjustment	64
Game Difficulty Scale, Competition Quality	Min. Rating	The minimum initial rating for the teams	1500
Game Difficulty Scale, Competition Quality	Max. Rating	The maximum initial rating for the teams	2500
Game Difficulty Scale, Competition Quality	Unknown Rating	The initial rating for teams whose average player values are unknown	1200
Recency Scale	α	The decay rate to de-emphasize older matches of the player exponentially	0.995
Ranking	d	Aggregated statistics definitions which are relevant for the analysis	User Provided
Preference Based Ranking	w	Preference weights provided by the ranker	User Provided
Objective Based Ranking	f_{obj}	Objective function provided by the ranker	User Provided

4.1.2 Parameters

The parameters of the proposed methodology are given in Table 7.

K parameter of the ELO algorithm is set to 64 to allow for rapid change in ratings. In [102], authors provide a discussion on the selection of K-factor and explain the effects of different parameter values. Based on their findings, parameter value of 64 represents higher uncertainty than average. Since the aim is to evaluate teams and players on a large scale, this uncertainty is a desired property. For recency scaling, α is chosen empirically. The α value provided in Table 7 corresponds to a scaling coefficient of 0.77 for a game that is a year old. The framework can be extended by defining further variables in the same fashion as described in Table 6.

Out of these parameters, team rating initialization values are based on the best practices coming from chess. In chess, 2500 is roughly assumed to be the grandmaster rating and 1200 is the lowest professional rating level [103].

Finally, user-defined analysis dimension parameter is the search space for the player performance. These are the definitions for variables for ranking. Definition of these cannot be avoided as the positions in Football have different requirements. The definitions specify how the in-game statistics will be aggregated and which variables from the dataset will be used. The full definitions for this experimental setup are given in Table 6 along with the user-defined weights w .

4.1.3 Models

The final rankings are computed via reducing the dimension ratings matrix to a final rating vector. The reduction approach is defined in detail in Section 3.3.3.

Optimisation parameters used in rankings towards an objective are given in Table 8.

Table 8: Parameters for Objective-Driven Rankings

Property	Value
tol_{max}	1e-7
$iter_{max}$	1000
$numBins$	10
f_{obj}	Change in Player Value
$trainSetSize$	70%
$validSetSize$	20%
$testSetSize$	10%

4.1.4 Metrics

To evaluate the performance of the proposed methodology, comparison to the existing methods in the literature is performed using relevant ranking metrics.

Most common and well established ranking evaluation metrics are Spearman’s Rank Correlation [104] and Kendall’s τ [105]. Both correlation coefficients quantify the overall agreement between two ranked arrays.

Correlation coefficients do not differentiate between correctly ranking top items vs. correctly ranking bottom items. In information retrieval problems, correctly ranking top items is more important than correctly ranking bottom ones [106]. Therefore, correlation coefficients are not adequate metrics for ranking problems with large items.

Average precision [107], which computes performance at several intervals in the rankings and reports the average, is a newer metric designed for information retrieval metrics. Average precision correlation is based on the assumption that correctly ranking the top items is more important. In the case where ranking system’s error is distributed randomly, average precision correlation is equivalent to Kendall’s τ . If the system has a higher precision for ranking the top-ranking items, then average precision correlation is higher than Kendall’s τ , and vice versa.

Normalised discounted cumulative gain [108] quantifies the gain obtained by the ranking system by correctly ranking items. In order to compute the gain, NDCG requires a gain vector for each of the ranks. Therefore it is not a static metric since the metric value depends on the specific importance of each rank.

Finally, a novel technique WS Coefficient [109] is shown to evaluate rankings in a more well-rounded way by giving more emphasis to correctly ranking top items. In addition, unlike NDCG, it does not require an input from users regarding the importance of correctly ranking observations.

	playerId	date	fee	mv	foot	team_rating
19	56	2017-07-01	27000000	18000000.00000	2.00000	2326.81905
20	56	2018-07-27	13500000	15000000.00000	2.00000	2263.82488
21	56	2020-10-05	11000000	12000000.00000	2.00000	2184.41393

Figure 12: Sample Rows from Dataset

4.2 Player Valuation

4.2.1 Dataset

Player valuation component of this thesis uses data from various sources online. The largest information on transfers and fees is provided by Transfermarkt [30].

A significant factor in player transfer fees is the player performance. This thesis combines the proposed ranking method and player valuation, therefore the raw player performance data from [29] is first transformed using the proposed methodology, then integrated into the dataset. In addition, team ratings used for the development of the scales is also integrated into player valuation modelling.

To quantify the impact of the player popularity, the search metric data for each player is collected from Google Trends. Player specific search trends are fused into the transfers and performance dataset using the date of the search metric and the id of the player.

The collected and fused dataset contains 70 thousand male football players from 85 countries and 174 competitions beginning at 2016-01-01. From Transfermarkt [30] 141121 transfers involving these players were collected, out of which 10680 were paid transfers with fee information. The rest of the transfers are free transfers, or loans. For the paid transfers, I computed the features as described in Section 3.4.3.

The final modelling dataset contains 220 features. The full set of features used in modelling is given in Appendix B with their definitions. The instances of the dataset are individual transfers linked to the player data such as performance and popularity. An example excerpt for Davy Klaassen is given in Figure 12. The excerpt contains "playerId" and "date" which are id variables to identify the transfer. The column "fee" is the target variable, and, "mv", "foot" and "team_rating" are sample predictors. The rest of the predictors are omitted from the image due to space concerns. First row in the image is Klaassen's transfer from Ajax to Everton, the second row is his transfer from Everton to Werder-Bremen and the final row contains relevant data for his transfer from Werder-Bremen to Ajax. All relevant features are computed up-to-date of the beginning of the transfer season of each transfer to avoid temporal data leakage.

The dataset has missing values where information is not available in a certain dimension for a player. For instance, players may not have a crowd-sourced market value or

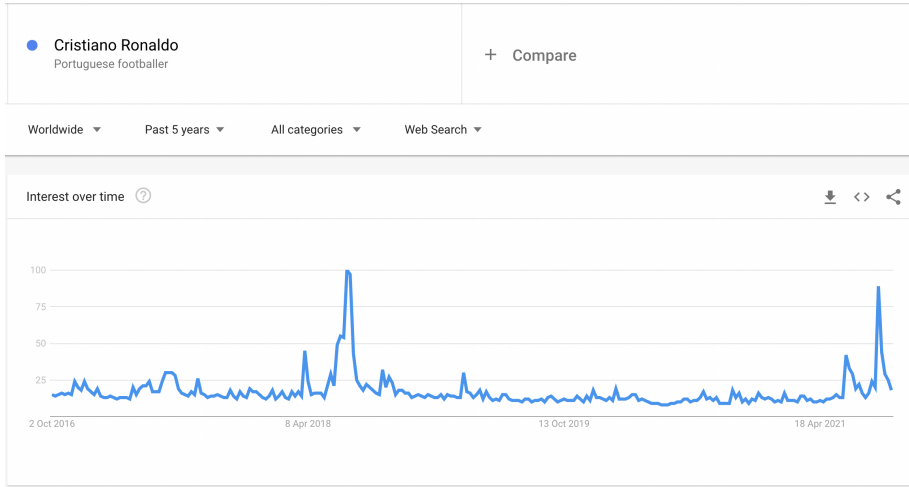


Figure 13: Search Term Frequency for Ronaldo

they may be in a youth team for which the team statistics are not available. Therefore, tree-based models, which can inherently deal with this type of problem, are chosen. The missingness statistics are given in Appendix C.

4.2.1.1 Popularity Data Collection

To quantify player popularity as a factor for modelling player transfer fees, the data was collected from Google Trends. The data collection was performed using the Google Trends REST API [90]. Google Trends data is a quantification of search frequency across a time-horizon. Multiple search terms can be compared within the same time horizon and the results are always scaled between [0-100], 100 point being the maximum frequency point for all selected search terms. Figure 13 and 14 show two search cases respectively. Figure 13 shows the frequency of search term 'Cristiano Ronaldo' within the past 5 years. Figure 14 shows the frequency of search terms 'Cristiano Ronaldo' and 'Football' in conjunction. When 'Football' is used in conjunction with 'Cristiano Ronaldo', Ronaldo's search frequency is scaled down significantly, even though he is one of the most popular players globally. When a player with limited international renown is searched, his relative frequency would diminish to 0. To compare different players fairly, a reference point is required that would work for most players. To compare players from various backgrounds, the relative frequency to their roles is used. Figures 15 and 16 respectively show Ronaldo's and Burak Yılmaz's relative popularity compared to their role. The ratio of this time series at the beginning of transfer season is used to quantify popularity.

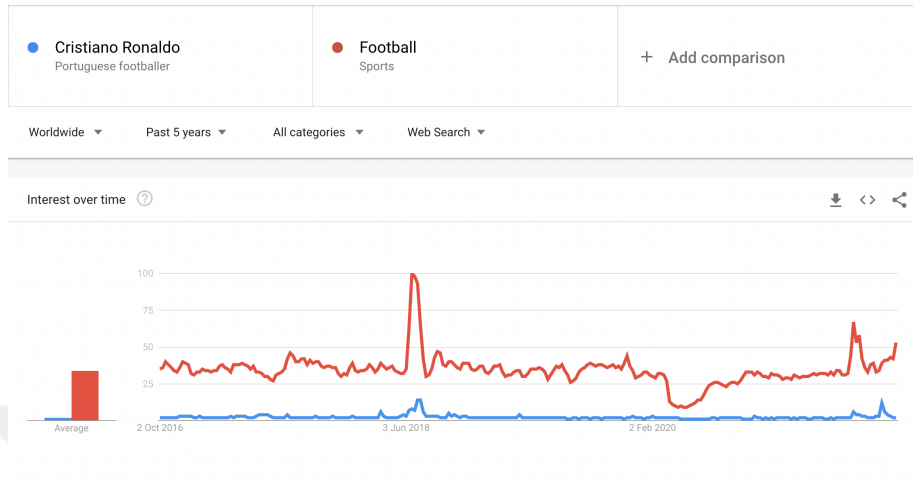


Figure 14: Search Term Frequency for Ronaldo and Football in Conjunction

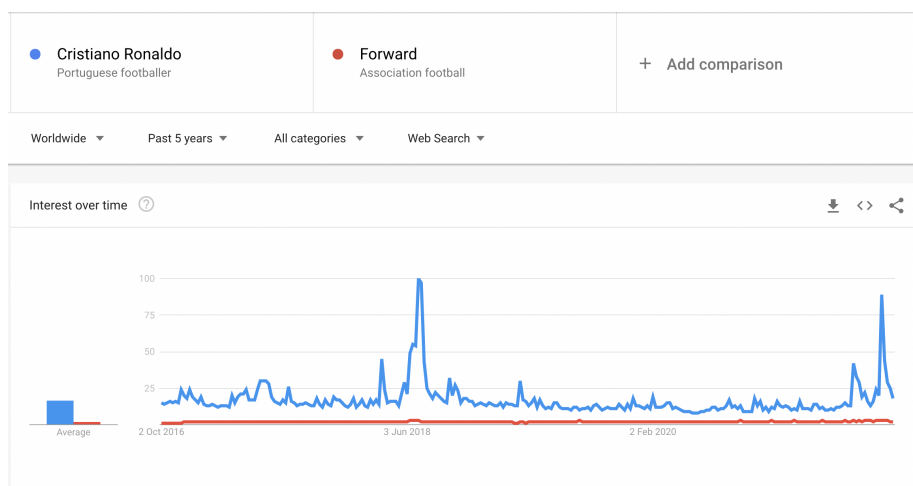


Figure 15: Search Term Frequency for Burak Yılmaz and Forward in Conjunction

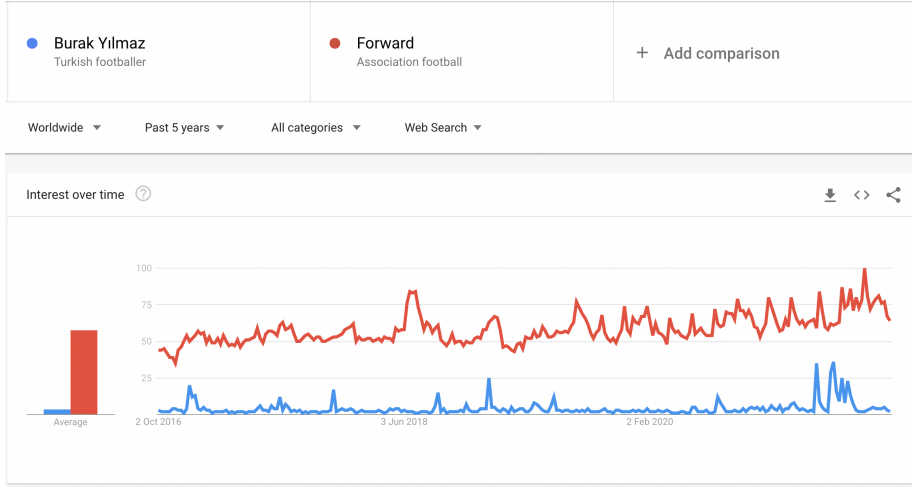


Figure 16: Search Term Frequency for Burak Yılmaz and Forward in Conjunction

Table 9: Ranking Setup for Player Valuation

Property	Value
Target Position	All
Minimum Play Duration	10 minutes
Minimum Number of Games	3
Ranking Method	Cosine PCA

4.2.2 Parameters

In case of player valuation, the goal is to estimate player transfer fees for all available players. Therefore the filters used in the ranking method are looser compared to the filters in Table 5. Ranking setup for player valuation is given in Table 9.

The dataset is split into train and hold-out set. Given the time-dependency of the problem the split must adhere to both out-of-time and out-of-sample splitting best practices. To evaluate performance fairly, all transfers in 2020 and onwards are put into the hold-out set. In addition, all transfers for 25% of unique players is reserved as held-out.

For model training, out-of-time cross validation with $K = 4$ is used. The dataset starts from 2015/2016 season with winter transfers and goes up to 2019/2020 season winter transfers (i.e only 2019 December transfers). Therefore, there are 5 separate seasons. The cross-validation scheme is shown in Figure 17.

To model the complex and large dataset, LightGBM [110] model is chosen as it is one of the most efficient algorithms to model mixed datasets. The hyperparameters of the model is optimized using Bayesian hyperparameter optimization using Optuna

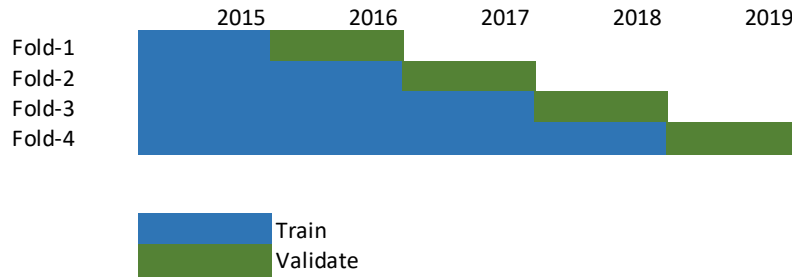


Figure 17: Cross-Validation Folds

Table 10: Hyperparameter Ranges

Hyperparameter	Minimum Value	Maximum Value
Learning Rate	5e-2	1e-1
Number of Estimators	10	500
Subsample	0.9	1
Minimum Child Samples	5	15
Maximum Depth	3	9
Column Sample by Tree	0.9	1
Number of Leaves	3	1000

[111]. The hyperparameter range is given in Table 10. The number of optimization trials is 100.

4.2.3 Models

This thesis builds 15 different machine learning models with different outcome variable treatments, thus different machine learning solutions. The discretisation of the target variable is performed using 5 bins, making equal frequency binning "quintile discretization".

- Single Models
 - Regression Models
 - * Using Un-transformed Fees
 - * Using Root-Transformed Fees
 - * Using Log-Transformed Fees
 - Ordinal Regression Models
 - * Quintile Discretization
 - Using Un-Transformed Fees

- * Natural Breaks Discretization
 - Using Un-Transformed Fees
 - Using Root-Transformed Fees
 - Using Log-Transformed Fees
 - Multi-Class Classification Models
 - * Quintile Discretization
 - Using Un-Transformed Fees
 - * Natural Breaks Discretization
 - Using Un-Transformed Fees
 - Using Root-Transformed Fees
 - Using Log-Transformed Fees
 - Multi-Label Classification Models
 - * Quintile Discretization
 - Using Un-Transformed Fees
 - * Natural Breaks Discretization
 - Using Un-Transformed Fees
 - Using Root-Transformed Fees
 - Using Log-Transformed Fees
- Model Ensembling
 - Stacking Using Lasso Regression
 - Stacking Using Decision Trees
 - Stacking Using Random Forest
 - Stacking Using LGBM

Notice that quintile discretisation is only applied to untransformed outcome variable. Quintiles perform sorting based frequency binning. Neither root transformation nor log transformation changes the order of the values. Therefore, quintiles of untransformed, root transformed and log transformed values would be exactly the same.

To extract further information from the various target variable transformations, the regression models are ensembled using stacking.

4.2.4 Metrics

Regression models output a continuous estimation. However, player transfer fees cannot be negative. To estimate the transfer fees using regression, this thesis uses "Truncated Mean Squared Error" (TMSE) defined in Eq. (13). Truncated MSE is chosen to compensate for the downside of regression modelling. Regression modelling performs estimation on an unbounded range. In case of player values, actual values are bounded in $[0, \infty)$. Therefore, negative predictions are truncated at 0 to

comply with the bounds of the outcome variable. Training metric is adjusted to reflect this domain-driven requirement.

$$TMSE(x, x_{true}) = \frac{\sum_{i=1}^N (\max(x, 0) - x_{true})^2}{N} \quad (13)$$

Unlike other discretized models, ordinal regression is also trained using TMSE. Whereas multi-class classification model is trained using logarithmic loss, which is a more suitable metric for skewed datasets. While equal-frequency binning results in equal-frequency classes, same cannot be said for Natural Breaks algorithm. To ensure the consistency across the classification methods, a single metric that is suitable for both is selected.

Finally, multi-label classification is trained using average precision score [107].



CHAPTER 5

RESULTS AND FINDINGS

To showcase the performance of the proposed framework, the methodology is applied to the experimental setup outlined in Section 4. The results are interpreted in isolation as well as in comparison to existing literature.

The studies in sports analytics are currently either small in scope or do not publish evaluation metrics. There is no standard set of metrics commonly accepted by researchers. Therefore the outputs of the proposed methodology are transformed in a comparable format to the existing studies where applicable, such as discretization of player rankings.

5.1 Player Performance Evaluation

Player performance evaluation results in an ordered output for the players based on their overall performance. Furthermore, the players are also ranked in each analysis dimension.

As stated in [1], there is no established form of validating player performances. The lack of a ground truth makes the validation non-trivial. To mitigate this, two dimensional validation for player ranks is adopted. First the players are ranked using the methodology, then following analyses are reported to provide a comprehensive view of the performance:

1. Based on the assumption that truly high-performing players would increase their market value after a high-performing period, this thesis compares the market valuations of players pre- and post- analysis. While market values depend on additional factors other than performance, the value change is expected to correlate positively with the player rankings.
2. Similarly, this thesis assumes that high-performing players would move to high-performing teams and vice versa. I perform team-fit validation by analysing the post-analysis ELO ratings of players' teams and show the rank correlation between the player rankings and team ELOs of players' destination teams in case of transfers. I only use the data for players who have been transferred to avoid data leakage. Since the team ELOs are used to calculate scales, the same

teams' ELO ratings cannot be used for validation. Therefore, the player should have moved to another team whose ELO rating is de-coupled from the scaled statistics.

3. For visual validation, players who are ranked in top-20 by the methodology are also provided as a separate list. In addition player market values and ages of the players at the end of 17/18 season are also provided for a more comprehensive overview.

The outputs of the methodology are compared to [1] which uses Composition of Probabilistic Preferences (CPP) using triangular distribution [112] to rank players *into groups of ranks (as opposed to assigning an individual rank to each player)* using the published R Package [113]. In addition, to compare to another MCDM approach, COMET [49] algorithm is applied to the analysis dimensions.

COMET method relies on having fuzzy membership numbers. To comply with the original application, I divided the standardized analysis dimensions into two distinct parts intervals (0-0.5), (0.5-1). Using these values for all 9 analysis dimensions, 19683 (3^9) characteristic objects were created. This method also relies on having expert feedback on deciding the superiority of one characteristic object over another to create Matrix of Expert Judgement (MEJ). However, in this study no such expert is available, therefore I use an expert decision function f_{exp} given in Eq. (14) to perform judgement where eps is an infinitesimally small number to address multiplication with 0. In this equation CO_i represents the i^{th} characteristic object and $\tilde{C}_{ji}$ is the fuzzy number for criteria (details are provided in [48]). This expert function allows to parallelize calculation of MEJ and removes the need for hierarchical modelling used in the original paper.

$$f_{exp}(CO_i) = \prod_{j=1}^r (\tilde{C}_{ji} + eps) \quad (14)$$

The ranking outputs are evaluated through search ranking evaluation metrics described in Section 4.1.4 because player ranking problem is analogous to ordering search results where there is no ground truth or relevance feedback available.

To compare the methodology internally, I report the metrics for all three proposed approaches. As a reminder, three approaches address three different use-cases of rankings. Method version 1 with preference weights supplied by the ranker detailed in Section 3.3.3.1, is aimed at addressing specific search requirements of the user. Method version 2 where the ranker does not know the exact search requirements but they are able to provide an objective they would like to achieve (Section 3.3.3.1) is aimed at ranking players based on their performance, as well as to optimise the given objective. Finally, where no information is available, method version 3 with Kernel-PCA (Section 3.3.3.3) is used to arrive at generally applicable performance rankings. In case of optimisation based version of the ranking, weight vector that maximizes

Table 11: Correlation Coefficients for Ranks Produced by Methodology

	Method Version 1	Method Version 2	Method Version 3
Method Version 1	1	0.999	0.928
Method Version 2	0.999	1	0.928
Method Version 3	0.928	0.928	1

the objective function value on the validation set is selected for further reporting as per Algorithm 1.

Table 11 shows the Spearman’s rank correlation coefficient matrix between the three ranking outputs of the proposed methodology. Based on these coefficients, it is clear that the methodology outputs internally consistent results. The preference weights supplied by the ranker and the optimisation approach are utilised for performing selection customisation where the performances are equivalent. For instance, given two players who have achieved the following ratings in dimensions A and B are numerically equivalent: $Player1 : \{A: 2, B: 1\}$ and $Player2 : \{A: 1, B: 2\}$. If the ranker values A more than B , then Player1 is ranked more highly than Player2, or, if the given objective is optimised by higher relative values of B then Player2 is ranked more highly. In case of method using Kernel-PCA, cosine transformation will map the players onto a new hyperspace relative to the rest of the dataset.

5.1.1 Financial Validation

In this section I provide financial evaluation metrics obtained from applying the three versions of the methodology. In addition, proposed methodology is compared to the existing work in the literature.

Figure 18 presents the cumulative investment gain by age group and ranks obtained through the methodology as well as through applying COMET [48] on the dataset. Return on investment is defined as acquiring the player *at the beginning* of the analysis period, and selling them after the analysis period ends.

The figure shows that, given the players are below a certain age, and they are ranked highly, the return on investment on high ranking players is positive. The higher the rank, the higher the return. For older players, the expected return is negative. This is due to players reaching the end of their careers, regardless of their ranking. For younger players, the return is positive regardless of their ranking, simply due to their initial values being relatively low and players having potential due to their age.

An interesting case is the middle age groups. There is a slight return on investment even the players are ranked low. This is due to inflation in player values across seasons.

The results show that version 3 of the methodology (i.e. Kernel-PCA version for no information case in Section 3.3.3.3) outperforms the second version of the methodology (Section 3.3.3.2), even though the objective function is the Spearman correlation between the final rankings and the magnitude of value change. There are several reasons for this. The first one is using the correlation instead of the magnitude of the value change as the objective function. The second one is the ranking computation. Cosine-Kernel PCA first maps the analysis dimensions onto a non-linear space, then rank aggregation is performed linearly. This non-linear mapping distributes the performance measurements in a smoother fashion, resulting in a smoother rank output.

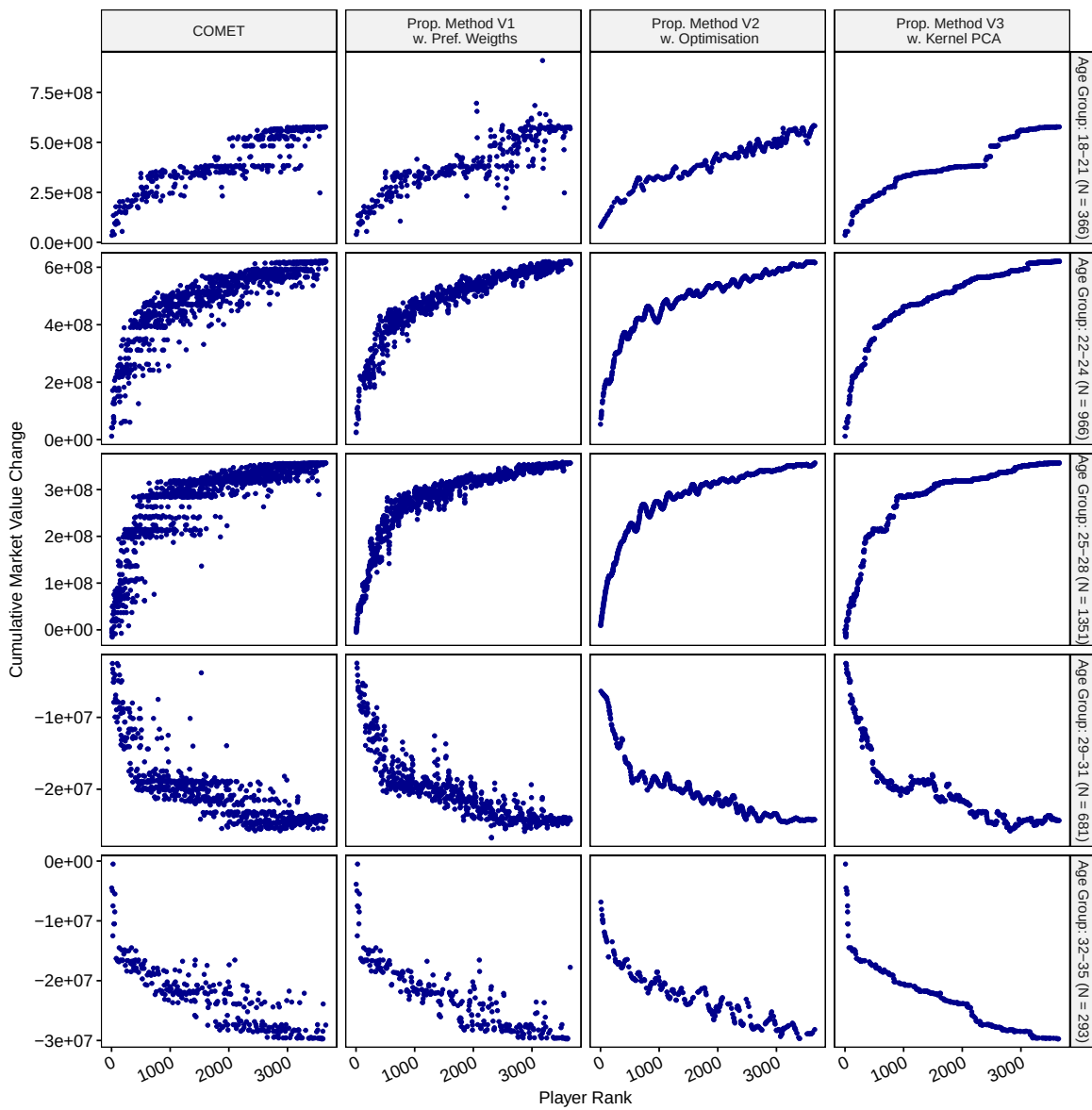


Figure 18: Cumulative Return on Investment per Age Group and Rank

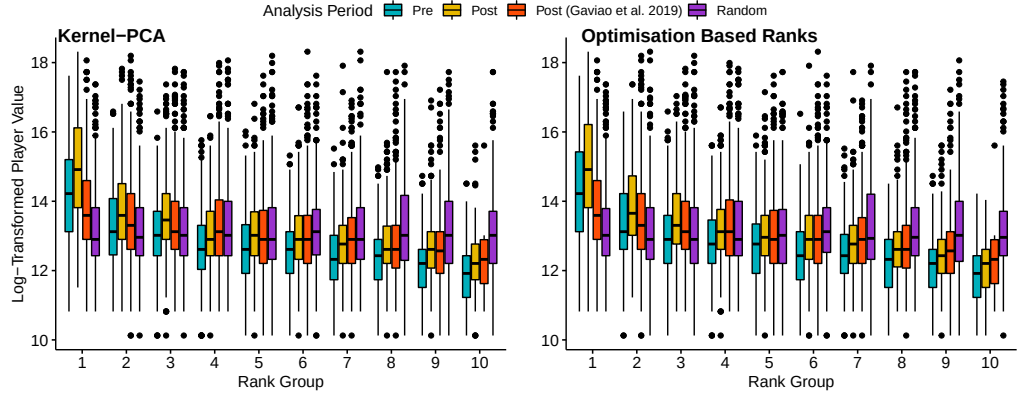


Figure 19: Player Market Value Distribution per Rank Group and Analysis Period

The proposed methodology outperforms COMET in terms of stability (dispersion of data points) of return on investment along ranks. One potential reason for this is the definitions of characteristic objects. More granular decision points could improve the stability of COMET method.

Regardless of the stability of the applied methodology, the ranking results reflect the player values changing from pre- analysis to post- analysis, proportionately to the rankings and age. The proposed statistical scaling and aggregation approach extract the performance information sufficiently for all the methods to perform adequate rankings.

To compare the results to CPP-Tri methodology [1], the ranks must be binned into groups as CPP-Tri methodology outputs groups of rankings instead of an ordered vector. The ratings produced by reducing the analysis dimensions into a single vector has more granular information than the ordered rank equivalent of the same vector. This rating vector is used to group the players into 10 distinct rank groups, Rank 1 being the best performers and Rank 10 being the worst. These rank groups are then compared in terms of the pre- and post- analysis value distributions. The rank groups of CPP-Tri are also added to comparison. Finally, post-analysis values of randomly assigned ranks are also provided for control. Only Kernel-PCA based ranks and optimisation based ranks are reported because the distributions of optimisation based ranks and preference based ranks are visually too similar to distinguish.

Figure 19 shows that the rank groups obtained through the proposed methodology have increased post-analysis values where rankings are high. In case of Rank 1 group, there is no overlap between the post-analysis values of the players and post-analysis values when ranks are randomly assigned. Interestingly, CPP-Tri method results in a ranking where the Rank 1 players have a lower median post-analysis value compared to their pre-analysis values. This is due to the age distribution of the players in rank groups. Figure 18 shows that with age, return on investment decreases, this explains the lower post-analysis values for the ranks obtained by CPP-Tri. Table 13 quantifies the ratings of the players in case of actual rank groups and random allocation.

Table 12: Two-Sample Kolmogorov-Smirnov Test P-Values

Rank Group	Version 3 vs. Random	Version 3 vs. Gaviao et. al	Version 2. vs. Gaviao et. al	Version 1 vs. Gaviao et. al	Gaviao et. al vs. Random
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0357	0.0002	0.0001	0.0047
3	0.0003	0.0034	0.0307	0.0354	1
4	0.5741	0.4399	0.2434	0.2440	1
5	1	0.6010	0.8215	0.7685	1
6	1	0.7928	0.8215	0.7685	1
7	1	0.7073	0.9708	0.9421	1
8	1	0.3848	0.8215	0.7685	1
9	1	0.7073	0.9708	0.9729	1
10	1	0.7928	0.9708	0.9421	1

For a more theoretically-based comparison, the results of Two-Sample Kolmogorov-Smirnov Statistical Test (one-tail) for comparing two non-parametric distributions are reported. Formally the hypothesis is defined as:

H_0 : Player values of the ranks calculated with the first methodology are drawn from the same distribution as the player values of the ranks calculated with the second methodology.

H_a : Player values of the ranks calculated with the first methodology are drawn from a distribution with a *greater* average than the distribution of the player values of the ranks calculated with the second methodology.

Table 12 shows the ‘False Discovery Rate’ corrected [114] p-values of the performed test, for ranks obtained via methodologies listed pairwise. The results show that both the methodology proposed in [1] have significantly greater player values than random allocation for first two rank groups at a significance level of 0.05. The statistical test shows that there is a significant evidence that our proposed methodology has greater player values in the third rank group in addition to the first two. Furthermore, the player values in the first rank group identified by our proposed methodology have significantly higher player values than the players identified as rank one by [1]. These results confirm that the proposed methodology is more successful in identifying players who will increase their market value beyond inflation based on their performance statistics. For the low rank groups (8^{th} – 10^{th}), statistical non-significance is expected because we perform a one-tailed hypothesis test (i.e. greater average).



Table 13: Rank Group Statistics Actual vs. Randomized

Rank Group	Rank Group						Randomized					
	Number of Players	Min. Rating	Max. Rating	Avg. Rating	Median Rating	Std. Rating	Number of Players	Min. Rating	Max. Rating	Avg. Rating	Median Rating	Std. Rating
1	347	1.5	2.17	1.63	1.61	0.11	334	0.7	1.84	1.21	1.22	0.22
2	331	1.39	1.5	1.44	1.44	0.03	325	0.63	1.97	1.18	1.17	0.23
3	347	1.31	1.39	1.35	1.35	0.02	346	0.68	1.88	1.19	1.19	0.22
4	360	1.25	1.30	1.28	1.28	0.02	331	0.63	1.98	1.2	1.19	0.22
5	344	1.19	1.25	1.22	1.22	0.02	328	0.65	1.95	1.19	1.17	0.22
6	370	1.13	1.19	1.16	1.16	0.02	378	0.69	1.83	1.21	1.21	0.22
7	305	1.08	1.13	1.10	1.10	0.01	327	0.49	2.17	1.2	1.18	0.25
8	341	1.01	1.08	1.05	1.05	0.02	361	0.54	1.91	1.19	1.19	0.24
9	366	0.91	1.00	0.96	0.97	0.03	402	0.65	1.91	1.21	1.20	0.23
10	346	0.49	0.91	0.83	0.84	0.07	325	0.58	1.91	1.21	1.19	0.22

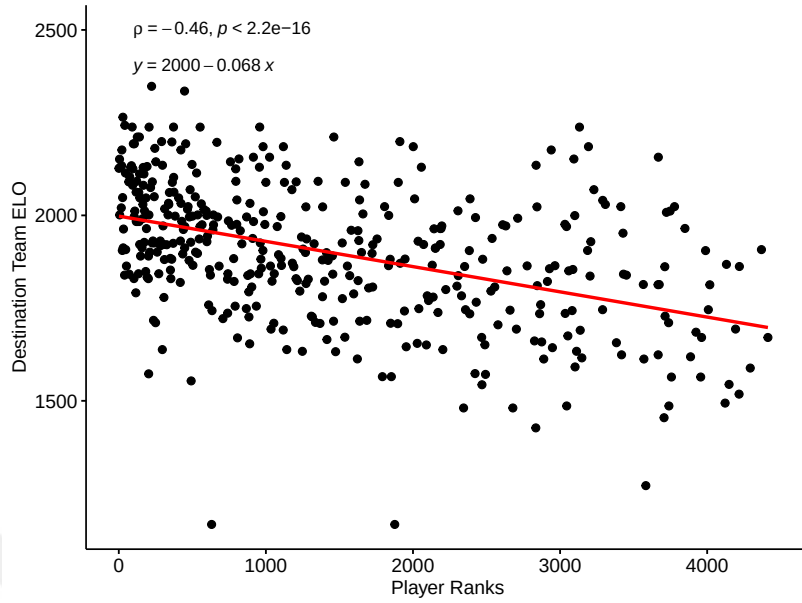


Figure 20: Relationship Between Ranks and Post-Analysis Destination Team Ratings

5.1.2 Team-Fit Validation

Team fit validation is based on the idea that high performing teams consist of high performing players. So player and team ranks are expected to have statistically significant correlation. However, team ratings are already utilized in creating of scales and aligning statistics for fair comparison. Therefore, to avoid data leakage, the validation can only be performed using the data of the players who have been transferred into a new team after the analysis. This prevents leakage from team ELO points into validation. In other words, by using the data from post-analysis transfers I decouple the scaling from output validation.

Figure 20 shows the relationship between the team player got transferred to (i.e. destination team) post-analysis and the identified ranking of the player. The figure also reports the Spearman rank correlation between the two values. The rank correlation is statistically significant in all confidence levels. I also report the linear regression least square estimation of the relationship between rankings and team ratings. Namely, the expected ELO of a player's post-analysis destination, decreases by 0.068 points with every step decrease in their rankings. Most ELO points are distributed between 1500 and 2000, which is in-line with Chess literature [53]. Team-fit validation is reported only for Kernel-PCA version as the linear-regression coefficients and the correlation is the same for all models.

5.1.3 Information Retrieval Metrics

Ranking players is an applied problem of ordering a set of items where there is no ground-truth available. Therefore, the correct rankings cannot be learned from feedback. Nor can the final output be validated by comparing to true rankings. This is a well-known problem in retrieval problems where ranking of search results are not customized to user preference (i.e. no feedback). The evaluation metrics developed for such problems are relevant for evaluating our results.

Similar to web search literature, in the absence of ground-truth, comparison of two-rankings is the established method of evaluating ranking outputs. As no other methodology outputs a comprehensive and comparable rankings, I opt to use post-analysis player values as a proxy to player performance. Note that similar to the return on player investment, these metrics cannot be computed for [1] since their approach does not produce an ordered list of players but rank groups instead.

To validate the results, I provide the following metrics:

1. Kendall's τ [105], which quantifies the overall agreement between two ranked arrays.
2. Average precision [107], which computes performance at several intervals in the rankings and reports the average.
3. Normalised discounted cumulative gain [108] which quantifies the gain obtained by the ranking system by correctly ranking items.
4. WS Coefficient [109] which is shown to evaluate rankings in a more well-rounded way by giving more emphasis to correctly ranking top items. In addition, unlike NDCG, it requires an input from users regarding the importance of correctly ranking observations.

In information retrieval problems, it is more important to rank the highest performers correctly than ranking lowest performers correctly [106]. Kendall's τ does not take this distinction into account. Conversely, average precision correlation is based on the assumption that correctly ranking the top items is more important. In the case where ranking system's error is distributed randomly, average precision correlation is equivalent to Kendall's τ . If the system has a higher precision for ranking the top-ranking items, then average precision correlation is higher than Kendall's τ , and vice versa. Furthermore, a novel technique WS coefficient [109] also gives more weight to top-ranked items.

Table 14 provides the results of the proposed method using the aforementioned metrics.



Table 14: The Results of the Proposed Solution using Information Retrieval Metrics

Metric	Version 3 w. Kernel-PCA	Version 2 w. Optimisation	Version 1 w. Pref. Weights	COMET score w. Proposed Scaling and Feat. Engineering	COMET score w. Raw In-Game Stats
Kendall's τ	0.364	0.407	0.407	0.307	-0.026
Average Precision 0.480 $\pm$ 0.070	0.601 $\pm$ 0.011	0.579 $\pm$ 0.040	0.606 $\pm$ 0.05	0.579 $\pm$ 0.040	0.549 $\pm$ 0.042
Normalised Cumulative Discounted Gain (NCDG)	0.707	0.692	0.692	0.704	0.672
WS Coefficient	0.987	0.975	0.975	0.845	0.262

All metrics show a positive correlation between the rankings provided by the methodology and the eventual player value. Furthermore, order-dependent metrics (i.e. metrics other than Kendall's τ) have higher values than τ , showing that the proposed methodology is able to capture the rankings of the highest-valued players.

In addition, COMET was applied to the proposed scaling and aggregation methodology behaves similarly to the results provided by the proposed methodology, however, the proposed methodology outperforms the results obtained with COMET. This is due to COMET's characteristic object representation dividing the feature space into buckets of preference, and treating all values fall into the same bucket with similar preference values. COMET applied to raw data performs significantly worse than the proposed methodology as well as COMET applied to scaled and aggregated data. This reflects the need for considering factors that affect performance into account while analyzing sports data globally.

5.1.4 Real-World Results

Current configuration of the methodology outputs the ranks for 3681 players, of which the top-20 players according to their ratings are listed in Table 15. This table shows the team and age at the beginning and end of the analysis period to provide information regarding the status of the top-ranked players. It also gives details on the player values obtained from Transfermarkt [30] right before and immediately after the analysis period.

Fourteen players in the top-20 have increased their market value at the end of analysis period. An exception to this as a young player is Luke Shaw. He sustained an injury which put him on the sidelines for 14 matches in season of 16/17. Shortly after that he sustained another injury that put him on the sidelines for another 11 weeks in the same season. This is potentially one reason why his market valuation has decreased immediately after the end of our analysis period. His current market value is 25M Euros. In other cases, market value decreases are explained by the age of the player which is a general pattern shown in Figure 18.

Figure 21 shows that all but one top-ranking players below age 25 have increased their market values after the period of analysis (i.e. after being identified as top-performers). The period for which the performance data was used in analysis is marked with a grey shade in the figure.

To enable the readers to test the methodology and the configurations, a public API¹, is provided which uses the publicly available data by [115]. I also provide a graphical user interface to compute the results of arbitrary configurations in the form of a dashboard² which uses the same API.

¹ <https://jss-api-aolebn4toq-ew.a.run.app/redoc>

² <https://jss-dashboard-aolebn4toq-ew.a.run.app>

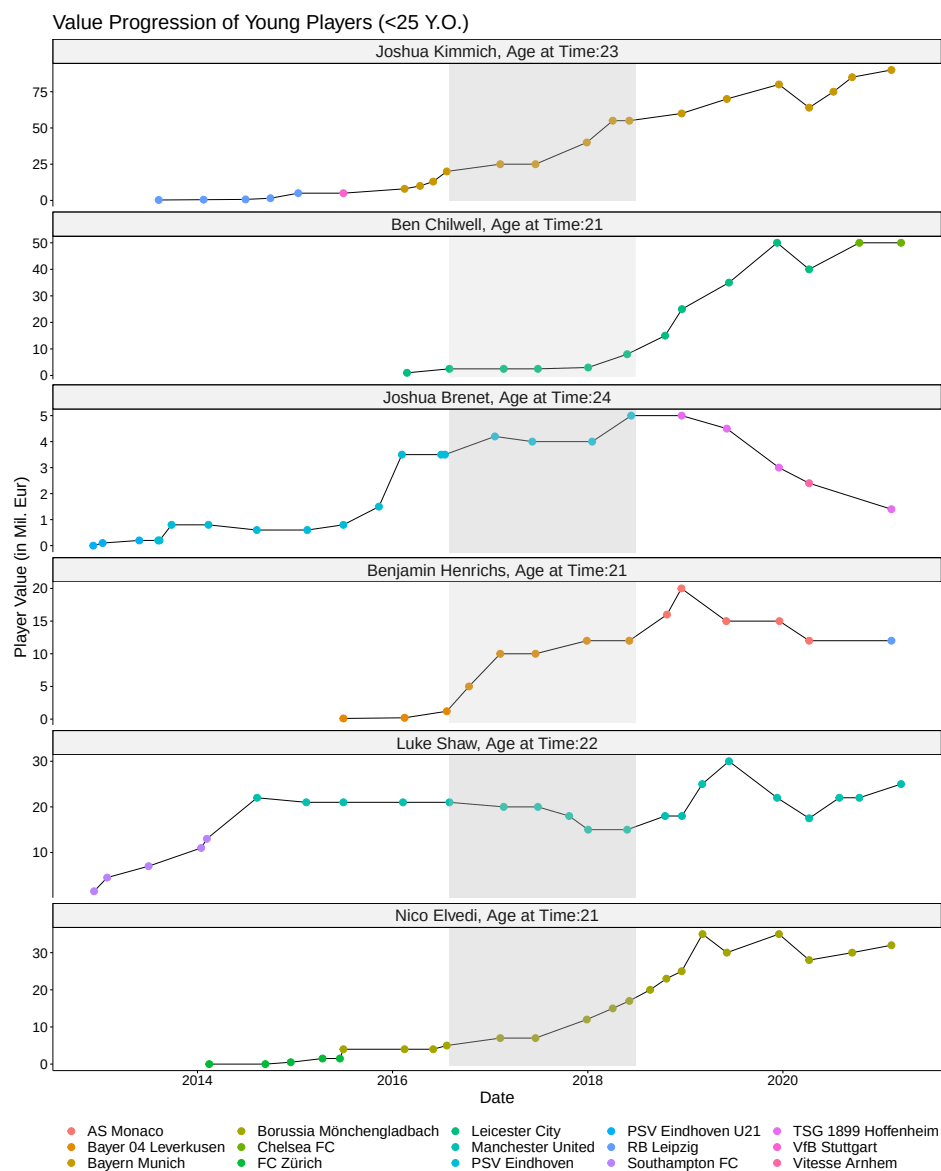


Figure 21: Market Value Changes Through Time



Table 15: Top 20 Players and Their Metadata

Rank	Name	Team Before Analysis	Team After Analysis	Market Value Before Analysis	Market Value After Analysis	Current League	Age at June, 2018
1	Fabian Delph	Manchester City	Manchester City	10M	15M	Premier League (England)	21
2	Luke Shaw	Manchester United	Manchester United	21M	18M	Premier League (England)	22
3	Joshua Kimmich	Bayern Munich	Bayern Munich	20M	60M	Bundesliga (Germany)	23
4	Ben Chilwell	Leicester City	Leicester City	2.5M	15M	Premier League (England)	21
5	Benjamin Henrichs	Bayer 04 Leverkusen	AS Monaco	1.25M	16M	League 1 (France)	21
6	Alessandro Florenzi	AS Roma	AS Roma	22M	30M	Seria A (Italy)	27
7	Kieran Gibbs	Arsenal FC	West Bromwich Albion	13M	6M	Premier League (England)	28
8	Davide Santon	Inter Milan	AS Roma	6M	5M	Seria A (Italy)	27
9	Christian Fuchs	Leicester City	Leicester City	6M	3.5M	Premier League (England)	32
10	Nico Elvedi	Borussia Mönchengladbach	Borussia Mönchengladbach	5M	20M	Bundesliga (Germany)	21
11	Francisco Femenia Far	Deportivo Alaves	Watford FC	1M	5M	Premier League (England)	27
12	Jonas Hector	1. FC Köln	1. FC Köln	14M	12.5M	Bundesliga (Germany)	28
13	Marcel Schmelzer	Borussia Dortmund	Borussia Dortmund	9M	5M	Bundesliga (Germany)	30
14	Sergio Roberto Carnicer	FC Barcelona	FC Barcelona	20M	55M	La Liga (Spain)	26
15	Joshua Brenet	PSV Eindhoven	TSG 1899 Hoffenheim	3.5M	5M	Bundesliga (Germany)	24
16	Íñigo Lekue Martínez	Athletic Bilbao	Athletic Bilbao	3M	4M	La Liga (Spain)	31
17	Kieran Trippier	Tottenham Hotspur	Tottenham Hotspur	6M	30M	Premier League (England)	27
18	Jeffrey Schlupp	Leicester City	Crystal Palace	8M	10M	Premier League (England)	25
19	Aleksandar Kolarov	Manchester City	AC Roma	11M	12M	Seria A (Italy)	32
20	David Olatukunbo Alaba	Bayern Munich	Bayern Munich	45M	55M	Bundesliga (Germany)	26

Table 16: Discretisation Statistics

Class	Y Transform	Y Discretisation	N	Avg Fee	Median Fee	Sd Fee	Max Fee	Min Fee
0	Log	Natural Breaks	10	€ -	€ -	€ -	€ -	€ -
1	Log	Natural Breaks	1077	€ 0.04	€ 0.05	€ 0.02	€ 0.09	€ 0.00
2	Log	Natural Breaks	2976	€ 0.30	€ 0.30	€ 0.15	€ 0.64	€ 0.10
3	Log	Natural Breaks	3049	€ 1.77	€ 1.50	€ 0.86	€ 3.90	€ 0.65
4	Log	Natural Breaks	1865	€ 13.41	€ 8.40	€ 14.87	€ 222.00	€ 3.94
0	None	Natural Breaks	8031	€ 1.43	€ 0.60	€ 1.81	€ 8.00	€ -
1	None	Natural Breaks	765	€ 14.67	€ 13.00	€ 5.04	€ 27.00	€ 8.10
2	None	Natural Breaks	158	€ 39.70	€ 36.50	€ 10.56	€ 68.00	€ 27.50
3	None	Natural Breaks	22	€ 96.59	€ 85.75	€ 23.69	€ 145.00	€ 70.00
4	None	Natural Breaks	1	€ 222.00	€ 222.00	€ -	€ 222.00	€ 222.00
0	Root	Natural Breaks	5044	€ 0.36	€ 0.28	€ 0.31	€ 1.10	€ -
1	Root	Natural Breaks	2497	€ 2.58	€ 2.30	€ 1.10	€ 5.10	€ 1.13
2	Root	Natural Breaks	974	€ 8.96	€ 8.00	€ 2.85	€ 15.50	€ 5.15
3	Root	Natural Breaks	406	€ 24.91	€ 22.45	€ 7.82	€ 45.30	€ 15.60
4	Root	Natural Breaks	56	€ 75.70	€ 64.38	€ 31.65	€ 222.00	€ 48.00
0	None	Quantile	1787	€ 0.07	€ 0.07	€ 0.05	€ 0.16	€ -
1	None	Quantile	1675	€ 0.30	€ 0.30	€ 0.08	€ 0.50	€ 0.17
2	None	Quantile	1901	€ 0.84	€ 0.80	€ 0.27	€ 1.48	€ 0.50
3	None	Quantile	1750	€ 2.35	€ 2.21	€ 0.66	€ 3.94	€ 1.50
4	None	Quantile	1864	€ 13.41	€ 8.44	€ 14.87	€ 222.00	€ 4.00

5.2 Player Valuation

Problem of estimating transfer fees in Football is used as a test-bed for measuring the performance of different machine learning modelling approaches. All models are built using LightGBM algorithm [110], but with different outcome variable representations. Hyperparameter search space is also the same for all models, however the final converged values are different in different models.

The models can be grouped in three categories: Continuous models, discrete models and stacked models. Continuous models estimate transfer fees with no discretisation, whereas discrete models first bin the target variable as described in Sections 3 and 4, then perform modelling.

The outputs of the models are compared to Transfermarkt to demonstrate industry standard performance. In addition, the final stacked model is compared to existing methods in the literature.

5.2.1 Discrete Models

This thesis compares various discrete models. There are different ways to discretise a continuous variable, as well as various classification approaches. Table 16 shows the results of discretisation of the target variable. In case of extreme values, natural breaks algorithm results in a single class with very few numbers. For instance, there are a few paid transfers in the dataset with a value of 1. These transfers are likely to be paid transfers where the transfer fee is unknown, but the data source (Transfermarkt) opted to assign a value to differentiate between the values of free transfers and paid transfers. Similarly, when data is not transformed, Kylian Mbappe's transfer of 222 Million € stands alone in a class.

Table 17: Classification Cross-Validation Metrics Across Years

Model	Transformation	Discretisation	Year	Kappa	Overall Accuracy	Macro F1
Multiclass Classification	Log	Natural Breaks	2016	0.7000	0.7892	0.6395
			2017	0.7005	0.7880	0.6367
			2018	0.6802	0.7738	0.6238
			2019	0.6845	0.7746	0.6163
	None	Natural Breaks	2016	0.7037	0.9468	0.5912
			2017	0.5605	0.9188	0.4085
			2018	0.6398	0.9334	0.4985
			2019	0.7234	0.9431	0.5135
		Quintile	2016	0.5988	0.6797	0.6846
			2017	0.6053	0.6856	0.6811
			2018	0.6083	0.6876	0.6898
			2019	0.6090	0.6876	0.6802
	Root	Natural Breaks	2016	0.6690	0.8022	0.6474
			2017	0.6677	0.7983	0.6248
			2018	0.7013	0.8222	0.7750
			2019	0.6916	0.8167	0.6606
Multilabel Classification	Log	Natural Breaks	2016	0.7079	0.7952	0.6439
			2017	0.6997	0.7880	0.6402
			2018	0.7018	0.7897	0.6344
			2019	0.6898	0.7786	0.6197
	None	Natural Breaks	2016	0.6808	0.9418	0.4232
			2017	0.6164	0.9267	0.4259
			2018	0.6870	0.9418	0.5601
			2019	0.7388	0.9472	0.5030
		Quintile	2016	0.6128	0.6908	0.7012
			2017	0.6209	0.6974	0.7035
			2018	0.6428	0.7148	0.7209
			2019	0.6397	0.7117	0.7093
	Root	Natural Breaks	2016	0.6796	0.8082	0.6676
			2017	0.6847	0.8077	0.6966
			2018	0.7021	0.8230	0.7420
			2019	0.7056	0.8241	0.6765
Ordinal Regression	Log	Natural Breaks	2016	0.6840	0.7791	0.6284
			2017	0.6909	0.7825	0.6332
			2018	0.6844	0.7776	0.6249
			2019	0.6910	0.7793	0.6199
	None	Natural Breaks	2016	0.7224	0.9488	0.4664
			2017	0.5923	0.9243	0.4303
			2018	0.6691	0.9395	0.4519
			2019	0.7179	0.9425	0.4914
		Quintile	2016	0.5590	0.6476	0.6575
			2017	0.6023	0.6822	0.6911
			2018	0.6067	0.6858	0.6961
			2019	0.6071	0.6852	0.6871
	Root	Natural Breaks	2016	0.6918	0.8143	0.6849
			2017	0.6745	0.8014	0.6219
			2018	0.6898	0.8147	0.7166
			2019	0.7155	0.8294	0.7544

Table 17 shows the cross-validation classification metrics across the models and years. As expected, in most cases performance increases with the year. This is expected because the 2016 fold is trained only using the data from 2015, while 2019 fold is trained using all the data up to it.

Table 18 shows the classification metrics on the test set. The skewness present in the player values is reflected in the model metrics. Natural breaks algorithm finds the most optimal separation point in a vector. Given skewed data, the resulting bins would have varying frequencies. In such cases, the resulting classes would be imbalanced. Accuracy of all models that are built using natural breaks applied to data without transformation are above 90%, yet the same models underperform in metrics that are more appropriate for imbalanced datasets. Root transformed models perform well overall across all metrics.

Table 18: Overall Classification Metrics

Model Type	Y Transform	Y Discretisation	Overall Accuracy	Macro F1	Kappa
Multiclass Classification	Log	Natural Breaks	0.75269733	0.60519536	0.65176978
Multiclass Classification	None	Quintile	0.66325951	0.66145164	0.57910734
Multiclass Classification	None	Natural Breaks	0.94378194	0.51435826	0.6466978
Multiclass Classification	Root	Natural Breaks	0.80011357	0.66745291	0.65120074
Multilabel Classification	Log	Natural Breaks	0.74190801	0.59618593	0.63543026
Multilabel Classification	None	Quintile	0.66865417	0.67417317	0.58591001
Multilabel Classification	None	Natural Breaks	0.939523	0.50172739	0.6178983
Multilabel Classification	Root	Natural Breaks	0.79613856	0.6507163	0.64168306
Ordinal Regression	Log	Natural Breaks	0.7552527	0.60456811	0.65283113
Ordinal Regression	None	Quintile	0.65445769	0.66182525	0.56822605
Ordinal Regression	None	Natural Breaks	0.94860875	0.57679594	0.68382962
Ordinal Regression	Root	Natural Breaks	0.81913685	0.7165859	0.68743183

Table 19: Balanced Accuracy by Class per Model

Model Type	Y Transform	Y Discretisation	Class 0	Class 1	Class 2	Class 3	Class 4
Multiclass Classification	Log	Natural Breaks	NA	0.8653	0.8287	0.7822	0.9032
Multiclass Classification	None	Quintile	0.8419	0.7972	0.7306	0.7299	0.8723
Multiclass Classification	None	Natural Breaks	0.9096	0.8409	0.8699	0.7496	NA
Multiclass Classification	Root	Natural Breaks	0.8863	0.7838	0.7963	0.8143	0.8203
Multilabel Classification	Log	Natural Breaks	NA	0.8673	0.8102	0.7737	0.9013
Multilabel Classification	None	Quintile	0.8924	0.7566	0.7316	0.7296	0.9021
Multilabel Classification	None	Natural Breaks	0.8994	0.8267	0.7987	0.7994	NA
Multilabel Classification	Root	Natural Breaks	0.8794	0.7765	0.8116	0.8402	0.7678
Ordinal Regression	Log	Natural Breaks	NA	0.9047	0.8286	0.7821	0.8907
Ordinal Regression	None	Quintile	0.9043	0.7501	0.7258	0.7151	0.9062
Ordinal Regression	None	Natural Breaks	0.9173	0.8492	0.8711	0.8996	NA
Ordinal Regression	Root	Natural Breaks	0.8936	0.8021	0.8106	0.8634	0.9080

Tables 19 and 20 respectively show the balanced accuracy and F1-Score by class per model. In case of classes with very few-instances, the test metrics are not available, these are shown as 'NA'.

5.2.2 Continuous Models

Three different continuous models are built for estimating the transfer fees, using the three transformations of the target variable. Out-of-time cross-validation errors of the three models, compared to the real world transfer fees are given in Figure 22.

Table 20: F1-Score by Class per Model

Model Type	Y Transform	Y Discretisation	Class 0	Class 1	Class 2	Class 3	Class 4
Multiclass Classification	Log	Natural Breaks	NA	0.7509	0.7621	0.7335	0.7795
Multiclass Classification	None	Quintile	0.7759	0.5844	0.5714	0.6033	0.7723
Multiclass Classification	None	Natural Breaks	0.9750	0.6409	0.4225	0.5333	NA
Multiclass Classification	Root	Natural Breaks	0.9031	0.6944	0.5855	0.5736	0.5806
Multilabel Classification	Log	Natural Breaks	NA	0.7405	0.7611	0.7111	0.7683
Multilabel Classification	None	Quintile	0.7988	0.6123	0.5893	0.5993	0.7712
Multilabel Classification	None	Natural Breaks	0.9731	0.6051	0.4304	0.5000	NA
Multilabel Classification	Root	Natural Breaks	0.8970	0.6784	0.6053	0.6063	0.4667
Ordinal Regression	Log	Natural Breaks	NA	0.7559	0.7867	0.7254	0.7549
Ordinal Regression	None	Quintile	0.7881	0.6329	0.5843	0.5693	0.7345
Ordinal Regression	None	Natural Breaks	0.9774	0.6616	0.5783	0.6667	NA
Ordinal Regression	Root	Natural Breaks	0.9094	0.7220	0.6420	0.6667	0.6429

Cross-Validation Error Distribution

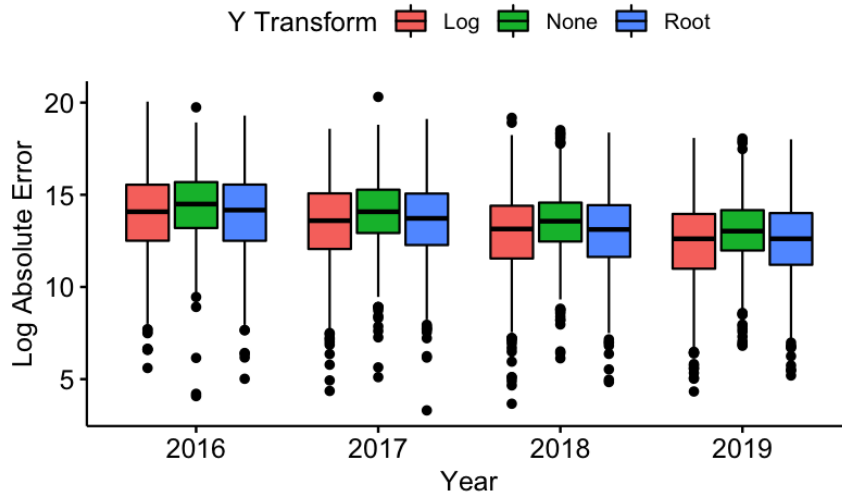


Figure 22: Regression Performance In CV Folds

The figure shows comparable median log-absolute error across the seasons, with a decrease from 2016 to 2019, as expected.

The regression model with no transformation consistently has higher error than the remaining two. The absolute errors are log-transformed for visualisation, so slight differences in the log absolute error translate to large absolute differences.

The performance metrics of different models on different levels of transfer fees are reported separately in Table 22. This table shows the test set Root Mean Squared Error (in million) of predictions made by base models, the final stacked prediction and predictions of TransferMarkt in decetile groups of player transfer fees. For clarity, the number of players in each transfer fee group is also reported. The table clearly shows that the log transform captures the transfers with low fee better while the root transformed model has overall the best predictive behavior of base models. The root transformed base learner also outperforms Transfermarkt predictions consistently. The base learner without transformation does not perform well in terms of RMSE, however as shown in Table 21, the base learner with no transform outperforms the rest of the predictions when the transfer fee is between 8.5 and 50 Million €.

Given the apparent inflation shown in Figure 18, the regression models are better suited to estimating transfer fees, as opposed to discretised predictions. Discretisation is employed when the grouping perform leads to better separation of the dataset, allowing the machine learning models to estimate average group values with better accuracy. In this case, model metrics are outperforming the industry standard in regression models, therefore, the inflation in player values is the dictating factor in model selection.

Table 21: Learner and TM Performance in Transfers Between 8.5 - 50 Million €(RMSE in Million, N = 300)

Learner	Error
Base Learner - No Transform	8.63
Base Learner - Root Transform	8.72
Base Learner - Log Transform	12.79
TM	9.33

Table 22: Prediction Performance on Test Set (RMSE in Million €)

Market Value Range	N	Base Learner No Transform	Base Learner Root Transform	Base Learner Log Transform	Stacked Model	TransferMarkt Predictions
$\leq 0.07M€$	345	0.43	0.25	0.15	0.06	0.26
$\leq 0.17M€$	338	0.62	0.44	0.38	0.70	0.55
$\leq 0.3M€$	382	0.80	0.52	0.41	0.29	0.52
$\leq 0.5M€$	362	0.86	0.63	0.47	0.24	0.96
$\leq 0.85M€$	276	1.12	0.84	0.66	1.11	0.97
$\leq 1.5M€$	412	1.70	1.27	1.05	0.58	1.62
$\leq 2.3M€$	278	1.67	1.48	1.30	1.62	1.40
$\leq 4M€$	349	2.17	1.99	2.19	2.36	2.52
$\leq 8.5M€$	319	3.59	3.34	3.85	3.11	4.05
$\leq 220M€$	317	11.12	10.79	16.59	8.65	10.38

Table 23: Meta-Learner Error Distribution on Test Set

Error Statistic	Transfermarkt	Lasso Regression	Decision Tree	Random Forest	LightGBM
Minimum Absolute Error (Mil. €)	0	0	0	0	0
Mean Absolute Error $\pm \sigma$ (Mil. €)	1.73 ± 4.07	1.14 ± 4.72	1.31 ± 5.5	1.16 ± 5.22	1.07 ± 4.73
Median Absolute Error (Mil. €)	0.5	0.54	0.27	0.29	0.26
Q3 Absolute Error (Mil. euro)	1.6	1.1	1.0	0.88	0.91
Maximum Absolute Error (Mil. €)	12.2	14.97	19.2	19.07	18.4

5.2.3 Stacked Models

To combine the information captured by the log-transformed regression model and root-transformed regression model, their outputs are ensembled using an additional meta-learners, which is also trained using the same hyperparameter optimization procedure. The meta-learners considered in the thesis are: Lasso Regression, Decision Tree Regression, Random Forests and LightGBM. Table 23 shows the performance of final models stacked using these meta-learners.

Table 23 shows that the best overall performing meta-learner is LightGBM. However, Transfermarkt’s crowd-sourced market value estimations are the closest predictions of high-value transfers, shown by the lowest maximum error. Random Forest, Decision Tree and LightGBM meta-learners perform similarly in terms of overall performance. Lasso regression meta-learner is better than tree-based learners in predicting high-value transfers shown by second lowest maximum error metric. This is due to linear regression being sensitive to outliers and being skewed by those values, regardless of regularisation.

To demonstrate the performance of the final model, the error distribution of final predictions on the test set compared to TransferMarkt predictions is given in Figure 23. Tables 24 and 25 show the real-world predictions, and demonstrate good predictions and bad ones respectively.

Finally, the relative importance of features for the top-20 features used by the meta-learner is given in Figure 24. The figure shows that the top features for final decision are the predictions from the base learners and Transfermarkt market value (Mv). The list of all features used, their definitions and families are provided in Appendix B.

5.2.4 Comparison With Existing Literature

Due to differences in data sources and data subsets, one-to-one comparison to existing literature is difficult. Therefore, for comparison, I applied methods in literature to the dataset in this thesis directly. The results are compared to [67], [74] and Lasso Regression Baseline in Table 27. I choose to compare to Lasso regression instead of the linear methods in the literature for descriptive modelling due to collinearity of the variables in the dataset and the resulting need for regularisation.

Table 24: Top-10 Closest Predictions Above 500,000 €

Name	Transfer Date	From	To	Transfer Fee	Proposed Method Predictions	TM Predictions
S. Haller	2021-01-08	West Ham	Ajax	22.5	23.5	30
M. Busi	2020-10-05	RSC Charleroi	Parma	7.5	7.68	2.5
A. Tchouaméni	2020-01-29	G. Bordeaux	Monaco	18	17.85	14
Maicon	2020-08-01	Galatasaray	Al-Nassr	1.43	1.48	4.8
O. Maritu	2020-02-17	SX Chang'an At.	SJZ Ever Bright	3.94	3.86	0.95
N. Okafor	2020-01-31	FC Basel	RB Salzburg	11.2	11.59	8
S. Sosa	2021-02-12	River Plate	Atlanta United	4.95	4.95	7.5
A. Zeqiri	2020-10-01	Lausanne-Sport	Brighton	4	4.03	1.5
L. Bittencourt	2020-07-15	TSG Hoffenheim	Werder Bremen	7	6.87	4.8
S. Ristovski	2021-02-02	Sporting CP	Dinamo Zagreb	1	0.96	3
S. Weissman	2020-08-31	Wolfsberger AC	Real Valladolid	4	4.06	6
Tao Qianglong	2020-02-27	HB CF	DL Pro Res.	2.61	2.51	0.6
Gilberto	2020-08-08	Fluminense	Benfica	3	2.85	1
L. Dykes	2020-08-19	Livingston FC	QPR	2.2	2.18	0.48
C. Domínguez	2020-08-24	Independiente	Austin FC	2.27	2.16	4
E. Rigoni	2021-05-26	Elche CF	São Paulo	1.8	1.72	3.2
A. Băluță	2020-07-31	Slavia Prague	Puskás AFC	0.7	0.71	2
Breno	2020-11-12	Juventude	Palmeiras	1.5	1.54	0.3
B. Kouyaté	2020-08-25	Troyes	FC Metz	3.5	3.33	2.2
R. Centurión	2020-07-24	Racing Club	Vélez Sarsfield	1.4	1.38	2.4

Table 25: Bottom-10 Worst Predictions Above 500,000 €

Player	Transfer Date	From	To	Transfer Fee	Proposed Method Predictions	TM Predictions
M. Pjanić	2020-09-01	Juventus	Barcelona	60	11.53	45
Arthur	2020-09-01	Barcelona	Juventus	72	24.21	56
T. Partey	2020-10-05	Atlético Madrid	Arsenal	50	11	40
V. Osimhen	2020-09-01	LOSC Lille	SSC Napoli	70	13.71	40
B. Chilwell	2020-08-26	Leicester	Chelsea	50.2	15.67	40
Álvaro Morata	2020-07-01	Chelsea	Atlético Madrid	35	11.51	36
D. van de Beek	2020-09-02	Ajax	Man Utd	39	12.79	44
Rúben Dias	2020-09-29	Benfica	Man City	68	14.42	35
J. David	2020-08-11	KAA Gent	LOSC Lille	27	9.09	25
Diogo Jota	2020-09-19	Wolves	Liverpool	44.7	12.69	28
P. Schick	2020-09-08	AS Roma	Bay. Leverkusen	26.5	11	25
Allan	2020-09-05	SSC Napoli	Everton	25	9.84	28
S. Sensi	2020-09-01	Sassuolo	Inter	20	7.9	20
A. Doucoure	2020-09-08	Watford	Everton	22.1	7.98	20
S. Dest	2020-10-01	Ajax	Barcelona	21	6.05	18
Gabriel	2020-09-01	LOSC Lille	Arsenal	26	8.08	20
D. Lovren	2020-08-01	Liverpool	Zenit S-Pb	12	0.28	12
T. Castagne	2020-09-03	Atalanta BC	Leicester	20	6.34	18
M. Doherty	2020-08-30	Wolves	Spurs	16.8	4.36	16
Lucas Paquetà	2020-09-30	AC Milan	Olympique Lyon	20	8.57	20

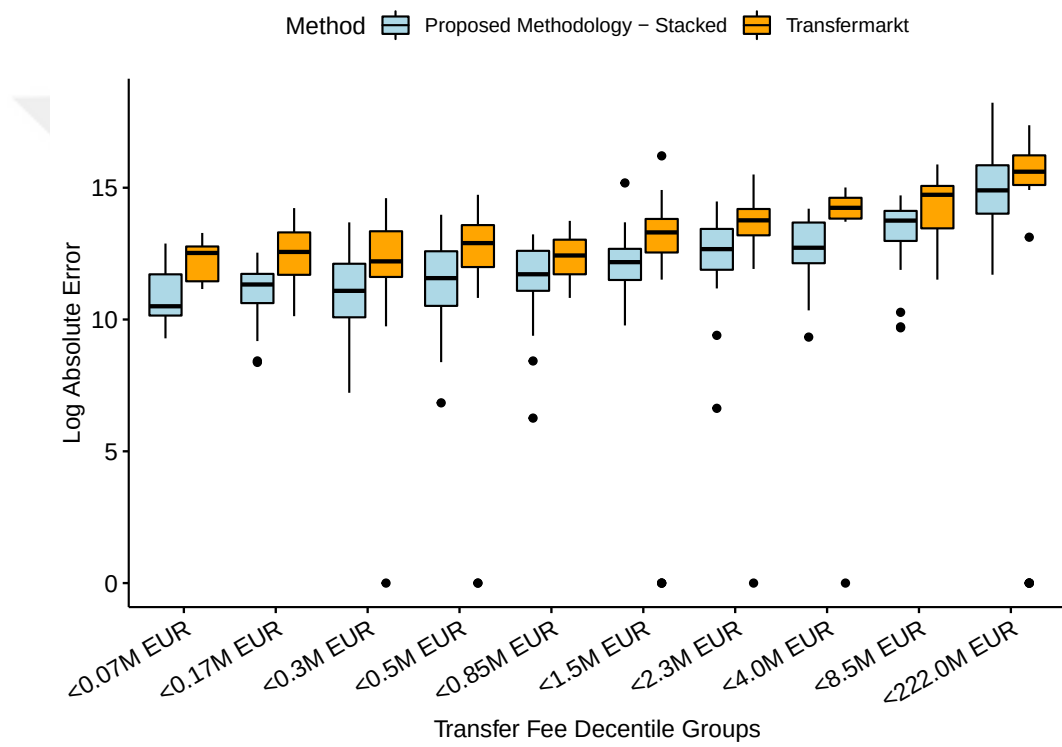


Figure 23: Error Distributions on Test Set Per Transfer Fee Buckets

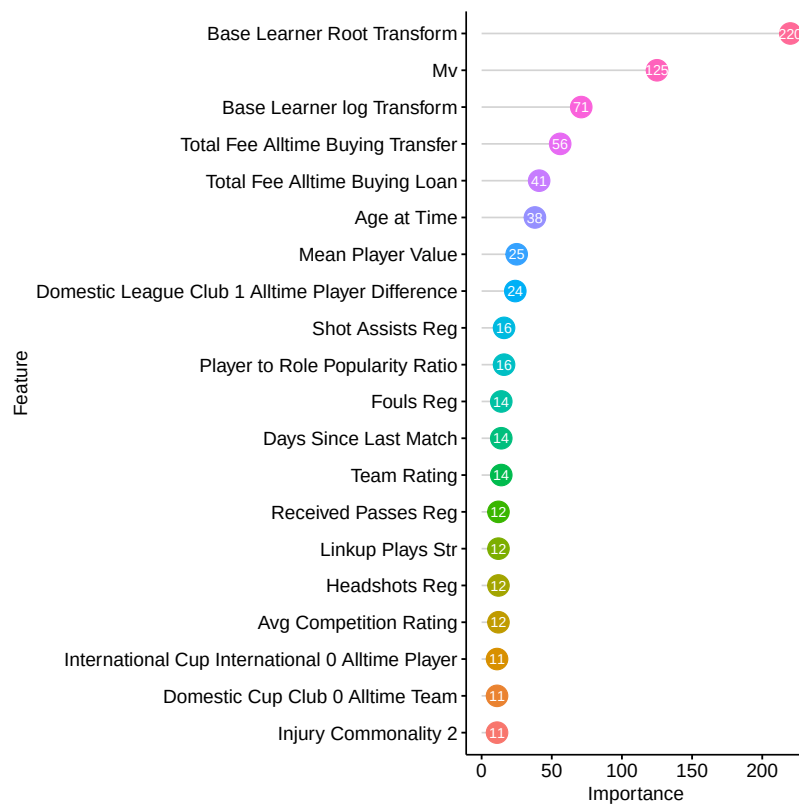


Figure 24: Most Important 20 Features

Table 26: POS Initial Parameter Values

Parameter	Value
c1	0.5
c2	0.5
w	0.9
k	30
p	2

Table 27: Comparison to Existing Literature

Method	% Error Change
Proposed Methodology	-6.12%
Lasso Regression Baseline	+6.89%
Muller et al. [67]	+3.6%
Yigit et. al [69]	-5.08%
Behravan et. al [74]	+23.85%

Same hyperparameter selection procedure is applied to the applied models. For Lasso regression baseline, alpha values in [0.1, 0.5] were considered. To compare with [74], the particle swarm optimisation hyperparameter optimization was initialized with parameters provided in Table 26 The number of particles to launch was held constant at 30.

The same dataset used in thesis was used to train these models. No missing value imputation was performed for the methods that can handle such datasets. For Lasso and POS-SVR, the missing values were imputed with -1 as this is out of the mathematical range of features with missing data.

Table 27 shows the % error change compared to TransferMarkt predictions and the sample size. For methods that perform well, error change is expected to be negative (i.e. decrease in error).

The results of [74] applied to the dataset of this thesis appear to be an outlier. The study implements a particle swarm optimisation to perform feature selection, followed by using Gaussian Kernel Support Vector Machines to perform regression on market values instead of transfer fees. Their dataset is small with fewer features. The kernel is used to map the data into a linearly-separable hyper-space, however this is not applicable to such a complex dataset as in this thesis, therefore the model diverges.



CHAPTER 6

CONCLUSION

This thesis provides a framework for analysing athlete performance globally in scope of Football. Athletic performance is further linked to transfer market dynamics, which can be complex and volatile.

While the proposed methodology is applied to Football, it can be extended to other team sports with similar global reach. In case of localized team sports, the proposed framework can still be applied by omitting various components.

Factors incorporated through scaling the in-game statistics allow for comparison between individual matches and individual players. This comparison was not possible before due to conditions that seemed uncontrollable or quantifiable up to this thesis. The alignment of in-game statistics enables a more granular assessment of players across their matches, as well as comparison of statistics of different players in different games. This means, following the framework, analysts can now pick arbitrary players and arbitrary matches and be able to compare performance by simply scaling the statistics as described, which was not possible before. Furthermore, devised scales and improved statistical quantification is shown to improve performance evaluation by other methods as well.

In addition, the proposed ranking framework various use-cases into account and outperforms existing methods in the literature in all cases. The existing ranking methods either require expert opinion or they cannot work with such large amounts of data. The proposed ranking methodology alleviates both restrictions while still allowing for experts to affect the decision making if need to be.

From a mathematical stand-point, a new logarithmic approximation to harmonic mean is used in the thesis. The approximation is empirically shown to have almost perfect alignment with harmonic mean calculations. Further research and experimentation is needed to verify that the arithmetic averages of logarithmic representation of individual measurements are equivalent to their harmonic mean, however the empirical results provided in the thesis is promising.

Finally, the thesis takes an information retrieval approach to athlete ranking problem which has not been applied previously in the literature. Sports analytics research can benefit from going beyond the established statistical analyses to cover more aspects of such a complex field.

The evaluation of the ranking methodology is a challenging topic due to lack of ground-truth and changing preferences/criteria. This thesis works around this issue by linking the general performance of players or player groups to the groups' financial and transfer performance. However, validating methodology on an individual player level proves difficult due to volatility in both individual player performance and player valuation in real world. Ideally, the change in rankings across time would be reflected in change in value or individual transfers, however due to market complexities this cannot be expected in real-life.

In terms of player valuation, to the best of our knowledge, the thesis provides the most comprehensive predictive modelling approach in the literature, combining proposed performance rankings, club and competition properties and search statistics of players. Volatility of the transfer fees provides grounds for applying advanced machine learning approaches. This thesis uses this complex problem to compare and contrast various machine learning representations and discuss their use-cases. The proposed player valuation modelling approach also outperforms the existing methods in the literature while using real-world data as opposed to synthetic data.

The use of skewed real-world data facilitated the application and comparison of various transformations to handle the skewness. While log transformation provides the most normalization of the data, in a use-case such as player transfers, it may be less than optimal when predicting outliers correctly is an important objective. Furthermore, when the target values exhibit a trend across time, discretization is not a suitable transformation for modelling, even though the model metrics may improve significantly.

This thesis provides an analysis into suitability to academically established non-linear transformations to real-world cases. In certain cases, soundest academic approach such as log transform may not be the best choice for the real-world use-case. In fact, often there is a trade-off between academic soundness and feasibility of real world solutions. To bridge this gap, there is a need for holding both academic studies and real-world solutions to a higher standard and approaching any machine learning problem from both perspectives.

As reported in [116], none of the thousands of machine learning publications regarding COVID-19 had a real-world application. This is also true for most ML studies. However, machine learning is primarily an applied science. Academic researchers in machine learning should aim to eventually facilitate meaningful, feasible and ethical real-world applications.

6.1 Contributions

This study has the following contributions to the literature:

1. A novel approach to reach per 90 in-game statistics, devoid of the effect of contextual factors, that approximates the established per 90 scaling averages across games and allows analysis on an individual game level.
2. A generic framework to evaluate players across competitions and a larger time frame, thus extending the pool of players. This framework is applicable beyond football, as it formally frames the impact of external factors as a scaling problem. The same framework can be applied to any global team sport.
3. A exponential decay approach to extend the time horizon for player performance evaluation.
4. Unsupervised discovery of rankings and an information retrieval perspective to player performance evaluation.
5. An optimization scheme to find the combination of skill-sets in a player to achieve an objective. This scheme is applied to discover the player performance profiles that maximize the return on investment, however it can also be applied to identify the profiles that contribute most to club placements or win likelihood. In other words, to identify key skills for sportive success
6. An expert-driven player search scheme to facilitate finding optimal profile players where preferences are known.
7. A data-driven player transfer fee estimation model that combines player roles, demographics, club and player performances, social media and player popularity metrics and player ranks that uses advanced machine learning techniques.
8. Application of advanced model ensembling techniques to combine predictions from various fee estimations to optimize the final predictions to different scenarios.
9. A formal comparison of various machine learning representations on modelling a skewed continuous variable
10. Predictive approach to modelling transfer fees using real-world data.

6.2 Limitations

This thesis provides means for reducing human input in rankings. However while doing so, it glosses over the requirements of the ranker gradually. This reduction in sensitivity to rankers' needs could be avoided if the gaps in the performance of the ranker team could be identified automatically. Studies have already been conducted towards automatic inference of team gaps and strategies on match logs [58], [59], [61], [117]. An extension of this thesis for the ranking step would be integration with such gap-identification approaches.

Another point where human intervention is needed is the definition of analysis dimensions. These must be defined by users or experts to quantify a certain aspect of

the game. Research into metric development could improve this aspect along with automatic identification of gaps.

Currently, the thesis uses the existing statistics and metrics provided by WyScout to further enrich them as data points. However, it is not guaranteed that these metrics are exhaustive or complete to represent player performance. Much of football, and other team sports that are played on a field, is about space-control much like strategy games. Currently, the available data (even events data) cannot provide this information because only a very small amount of time is spent on on-the-ball actions by players. Further research into using events data and soon-to-be-available position data to develop more meaningful metrics would contribute significantly to results and applicability of analytics in football.

Transfer fees are only a part of the transfer market. In addition to paid transfers, there are also free transfers, loans and internal promotions from youth teams to adult competitions. To have a more complete understanding of the financial landscape, the most likely type of transfer for a given player is also an important piece of information. Lack of this information is one of the biggest limitations of this thesis.

A general limitation that is out of control of the researcher of the thesis is the lack of mature applications and research in the field. As an emerging field, sports analytics have few studies where evaluation metrics and results have been reported. This limits the number of studies the proposed framework is compared to. Existing methods are also limited to a small subset of available data, therefore, even the studies where evaluation metrics are reported are incomparable.

6.3 Future Research Directions

To extend this thesis, there are two immediate research directions. The first one is research towards automatic identification of gaps of athletic performance and translating these gaps into high-level statistical aggregations to be used in ranking. While there is research on gap identification, this is not yet linked to metric development. The potential for streamlining analysis is still unknown. This could be an interesting prospect for future research and it is my opinion that, if successful, these efforts could revolutionize the sports analytics field.

Second research direction is adapting the proposed framework to other team sports. For instance, European players that get transferred to NBA seldom perform as well. There is a clear gap between either play styles or sheer quality of European basketball and NBA. The proposed framework can be applied to Basketball players to quantify the gap and identify the players that can achieve success in NBA.

REFERENCES

- [1] L. O. Gavião, A. P. Sant’Anna, G. B. A. Lima, and P. A. d. A. Garcia, “Evaluation of soccer players under the Moneyball concept”, *Journal of Sports Sciences*, vol. 0, no. 0, pp. 1–27, Dec. 2019.
- [2] FIFA. (Jul. 2021). “FIFA Annual Report 2020”, [Online]. Available: <https://publications.fifa.com/en/annual-report-2020/2020-financials-and-2022-budget/2020-financials-in-review/> (visited on 2021-07-22).
- [3] (). “Annual Review of Football Finance 2020 | Deloitte UK”, [Online]. Available: <https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html>.
- [4] S. Hoey, T. Peeters, and F. Principe, “The transfer system in European football: A pro-competitive no-poaching agreement?”, *International Journal of Industrial Organization*, vol. 75, p. 102 695, Mar. 2021.
- [5] E. Franck, “Financial Fair Play in European Club Football: What Is It All About?”, *International Journal of Sport Finance*, vol. 9, no. 3, pp. 193–217, 2014, Publisher: Fitness Information Technology.
- [6] E. Galariotis, C. Germain, and C. Zopounidis, “A combined methodology for the concurrent evaluation of the business, financial and sports performance of football clubs: the case of France”, *Annals of Operations Research*, vol. 266, no. 1, pp. 589–612, Jul. 2018.
- [7] C. Lucifora and R. Simmons, “Superstar Effects in Sport: Evidence From Italian Soccer”, *Journal of Sports Economics*, vol. 4, no. 1, pp. 35–55, Feb. 2003.
- [8] E. v. d. Berg, “The valuation of human capital in the football player transfer market”, M.S. thesis, Aug. 2011.
- [9] B. Gerrard, “Achieving transactional efficiency in professional team sports: the theory and practice of player valuation”, *Handbook on the Economics of Professional Football*, Nov. 2014, ISBN: 9781781003176 Publisher: Edward Elgar Publishing Section: Handbook on the Economics of Professional Football.
- [10] C. J. Anderson, L. Arrondel, A. Blais, J.-F. Daoust, J.-F. Laslier, and K. Van der Straeten, “Messi, ronaldo, and the politics of celebrity elections: voting for the best soccer player in the world”, *Perspectives on Politics*, vol. 18, no. 1, 91–110, 2020.

- [11] D. S. Sergio, M. Raul, and S. Eliza. (2013). “Hidden power law patterns in the top European football leagues”, [Online]. Available: <https://mp.ra.ub.uni-muenchen.de/48226/>.
- [12] W. B. Arthur, “Complexity economics”, *Complexity and the Economy*, 2013.
- [13] D. Barron, G. Ball, M. Robins, and C. Sunderland, “Artificial neural networks and player recruitment in professional soccer”, *PLOS ONE*, vol. 13, no. 10, e0205818, Oct. 2018.
- [14] D. Barron, G. Ball, M. Robins, and C. Sunderland, “Identifying playing talent in professional football using artificial neural networks”, *Journal of Sports Sciences*, vol. 0, no. 0, pp. 1–10, Jan. 2020.
- [15] StatsBomb. (Aug. 2013). “An Introduction To The ‘Per 90’ Metric”, [Online]. Available: <https://statsbomb.com/2013/08/an-introduction-to-the-per-90-metric/> (visited on 2021-03-30).
- [16] C. P. Barros and S. Leach, “Analyzing the Performance of the English F.A. Premier League With an Econometric Frontier Model:” *Journal of Sports Economics*, Aug. 2016.
- [17] S. Dobson and B. Gerrard, “The Determination of Player Transfer Fees in English Professional Soccer”, *Journal of Sport Management*, vol. 13, no. 4, pp. 259–279, Oct. 1999, Publisher: Human Kinetics, Inc. Section: Journal of Sport Management.
- [18] H. Kedar-Levy and M. Bar-Eli, “The Valuation of Athletes as Risky Investments: A Theoretical Model”, *Journal of Sport Management*, vol. 22, no. 1, pp. 50–81, Jan. 2008, Publisher: Human Kinetics, Inc. Section: Journal of Sport Management.
- [19] E. Nsolo, P. Lambrix, and N. Carlsson, “Player valuation in European football (Extended version)”, in *International Workshop on Machine Learning and Data Mining for Sports Analytics*, Springer, 2018, pp. 42–54.
- [20] C. Reep and B. Benjamin, “Skill and chance in association football”, *Journal of the Royal Statistical Society. Series A (General)*, vol. 131, no. 4, pp. 581–585, 1968, ISBN: 0035-9238 Publisher: JSTOR.
- [21] M. Lewis, *Moneyball: The Art of Winning an Unfair Game*. WW Norton & Company, 2004.
- [22] D. S. Mason, “Moneyball as a supervening necessity for the adoption of player tracking technology in professional hockey.”, *International Journal of Sports Marketing & Sponsorship*, vol. 8, no. 1, 2006.
- [23] M. Hughes, T. Caudrelier, N. James, I. Donnelly, A. Kirkbride, and C. Duschesne, “Moneyball and soccer: An analysis of the key performance indicators of elite male soccer players by position”, *Journal of Human Sport and Exercise*, vol. 7, no. 2, pp. 402–412, Jun. 2012.

- [24] D. Weimar and P. Wicker, “Moneyball revisited: Effort and team performance in professional soccer”, *Journal of Sports Economics*, vol. 18, no. 2, pp. 140–161, 2017.
- [25] M. Stewart, H. Mitchell, and C. Stavros, “Moneyball applied: Econometrics and the identification and recruitment of elite Australian footballers”, *International Journal of Sport Finance*, vol. 2, no. 4, pp. 231–248, 2007.
- [26] InStat. (2020). “InStat”, [Online]. Available: <https://football.instatscout.com/> (visited on 2020-11-12).
- [27] S. StatsBomb. (2020). “Where art thou german centre backs?”, [Online]. Available: <https://statsbomb.com/2019/11/where-art-thou-german-centre-backs/> (visited on 2020-11-12).
- [28] O. Opta. (2020). “Opta Sports”, [Online]. Available: <https://www.optasports.com/sports/football/> (visited on 2020-11-12).
- [29] W. Wyscout. (2020). “Wyscout”. en-US, [Online]. Available: <https://wyscout.com/> (visited on 2020-11-12).
- [30] T. Transfermarkt. (2020). “Transfermarkt”, [Online]. Available: <https://www.transfermarkt.com/> (visited on 2020-11-12).
- [31] UEFA. (2020). “UEFA Coefficients”, [Online]. Available: <https://www.uefa.com/memberassociations/uefarankings/> (visited on 2020-11-12).
- [32] O. Opta. (May 2020). “WhoScored.com”, [Online]. Available: <https://www.whoscored.com/>.
- [33] O. Opta. (2020). “The inner beauty of the beautiful game: How Opta analyses football”, [Online]. Available: <https://www.stuff.tv/features/inner-beauty-beautiful-game-how-opta-analyses-football>.
- [34] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, “PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach”, *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 5, pp. 1–27, Nov. 2019. arXiv: 1802.04987.
- [35] StatsBomb. (May 2012). “Football like never before”, [Online]. Available: <https://statsbomb.com/> (visited on 2021-03-30).
- [36] J. A. Martinez and L. Martínez, “A stakeholder assessment of basketball player evaluation metrics”, *Journal of Human Sport and Exercise*, vol. 6, no. 1, pp. 153–183, Mar. 2011.
- [37] S. Pettigrew, “Assessing the offensive productivity of NHL players using in-game win probabilities”, in *9th Annual MIT Sloan Sports Analytics Conference*, vol. 2, 2015, p. 8.
- [38] O. Schulte, Z. Zhao, M. Javan, and P. Desaulniers, “Apples-to-apples: Clustering and ranking NHL players using location information and scoring impact”, in *Proceedings of the MIT Sloan Sports Analytics Conference*, 2017.

- [39] S. McIntosh, S. Kovalchik, and S. Robertson, “Validation of the Australian football league player ratings”, *International Journal of Sports Science & Coaching*, vol. 13, no. 6, pp. 1064–1071, 2018.
- [40] S. Pradhan, “Ranking regular seasons in the NBA’s Modern Era using grey relational analysis”, *Journal of Sports Analytics*, vol. 4, no. 1, pp. 31–63, 2018.
- [41] J. Duch, J. S. Waitzman, and L. A. N. Amaral, “Quantifying the Performance of Individual Players in a Team Activity”, *PLOS ONE*, vol. 5, no. 6, e10937, Jun. 2010.
- [42] I. G. McHale, P. A. Scarf, and D. E. Folker, “On the Development of a Soccer Player Performance Rating System for the English Premier League”, *Interfaces*, vol. 42, no. 4, pp. 339–351, Jul. 2012.
- [43] J. Brooks, M. Kerr, and J. Gutttag, “Developing a data-driven player ranking in soccer using predictive model weights”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 49–55.
- [44] P. Power, H. Ruiz, X. Wei, and P. Lucey, “Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data”, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1605–1613.
- [45] A. Cakmak, A. Uzun, and E. Delibas, “Computational modeling of pass effectiveness in soccer”, *Advances in Complex Systems*, vol. 21, no. 03n04, p. 1 850 010, 2018.
- [46] G. Liu, Y. Luo, O. Schulte, and T. Kharrat, “Deep soccer analytics: learning an action-value function for evaluating soccer players”, *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1531–1559, 2020.
- [47] T. Pereira, J. Ribeiro, F. Grilo, and D. Barreira, “The Golden Index: A classification system for player performance in football attacking plays”, *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, vol. 233, no. 4, pp. 467–477, 2019.
- [48] W. Sałabun, A. Shekhovtsov, D. Pamučar, J. Wątróbski, B. Kizielewicz, J. Więckowski, D. Bozanić, K. Urbaniak, and B. Nyczaj, “A Fuzzy Inference System for Players Evaluation in Multi-Player Sports: The Football Study Case”, *Symmetry*, vol. 12, no. 12, p. 2029, Dec. 2020, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
- [49] S. Faizi, T. Rashid, W. Sałabun, S. Zafar, and J. Wątróbski, “Decision Making with Uncertainty Using Hesitant Fuzzy Sets”, *International Journal of Fuzzy Systems*, vol. 20, no. 1, pp. 93–103, Jan. 2018.
- [50] V. Darley, “Emergent Phenomena and Complexity”, in *Artificial Life IV. Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, MIT Press, 1994, pp. 411–416.

- [51] E. Triantaphyllou, “Multi-criteria decision making methods”, in *Multi-criteria decision making methods: A comparative study*, Springer, 2000, pp. 5–21.
- [52] L. M. Hvattum and H. Arntzen, “Using ELO ratings for match result prediction in association football”, *International Journal of Forecasting*, Sports Forecasting, vol. 26, no. 3, pp. 460–470, Jul. 2010.
- [53] A. E. Elo, *The Rating of Chessplayers, Past and Present*. Arco Pub., 1978.
- [54] F. Wunderlich and D. Memmert, “The Betting Odds Rating System: Using soccer forecasts to forecast soccer”, *PLOS ONE*, vol. 13, no. 6, e0198668, Jun. 2018.
- [55] J. Ali, S. Shahram, and M. Thomas, “Modeling team compatibility factors using a semi-Markov decision process: a data-driven approach to player selection in soccer”, *Journal of Quantitative Analysis in Sports*, vol. 9, no. 4, pp. 347–366, 2013.
- [56] L. Pappalardo and P. Cintia, “Quantifying the relation between performance and success in soccer”, *Advances in Complex Systems*, vol. 21, no. 03n04, p. 1 750 014, 2018.
- [57] K. Palczewski and W. Sałabun, “Identification of the football teams assessment model using the COMET method”, *Procedia Computer Science*, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019, vol. 159, pp. 2491–2501, Jan. 2019.
- [58] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, “Assessing team strategy using spatiotemporal data”, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 1366–1374.
- [59] Q. Wang, H. Zhu, W. Hu, Z. Shen, and Y. Yao, “Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications”, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 2197–2206.
- [60] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [61] T. Decroos, J. Van Haaren, and J. Davis, “Automatic discovery of tactics in spatio-temporal soccer match data”, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 223–232.
- [62] S. Green. (2012). “Assessing the Performance of Premier League Goalscorers”, [Online]. Available: <https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/> (visited on 2021-07-23).
- [63] C. Biermann, *Football Hackers: The Science and Art of a Data Revolution*. Bonnier Books Limited, 2019.

- [64] D. University. (Jul. 2021). “Metrics – The Soccer Analytics Revolution”. en-US, [Online]. Available: <https://sites.duke.edu/socceranalyticsrevolution/metrics/> (visited on 2021-07-23).
- [65] M. Thompson. (Jan. 2021). “The 5 key football metrics to know about in 2021”. en-GB, [Online]. Available: <https://www.twenty3.sport/data-analytic-s-5-key-football-metrics-to-know-about-in-2021/> (visited on 2021-07-23).
- [66] S. Herm, H.-M. Callsen-Bracker, and H. Kreis, “When the crowd evaluates soccer players’ market values: Accuracy and evaluation attributes of an online community”, *Sport Management Review*, vol. 17, no. 4, pp. 484–492, Nov. 2014.
- [67] O. Müller, A. Simons, and M. Weinmann, “Beyond crowd judgments: Data-driven estimation of market value in association football”, *European Journal of Operational Research*, vol. 263, no. 2, pp. 611–624, Dec. 2017.
- [68] P. Papadimitriou and H. Garcia-Molina, “Data leakage detection”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 51–63, 2010, ISBN: 1041-4347 Publisher: IEEE.
- [69] A. T. Yiğit, B. Samak, and T. Kaya, “Football Player Value Assessment Using Machine Learning Techniques”, in *Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making*, ser. Advances in Intelligent Systems and Computing, Cham: Springer International Publishing, 2020, pp. 289–297.
- [70] A. Yigit, B. Samak, and T. Kaya, “An XGBoost-lasso ensemble modeling approach to football player value assessment”, *J. Intell. Fuzzy Syst.*, 2020.
- [71] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system”, in *Proceedings of the 22nd Acm SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [72] R. Tibshirani, “Regression shrinkage and selection via the lasso”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, ISBN: 0035-9246 Publisher: Wiley Online Library.
- [73] S. Interactive. (Jul. 2021). “Football Manager - Create Your Football Story”, [Online]. Available: <https://footballmanager.com/> (visited on 2021-07-23).
- [74] I. Behravan and S. M. Razavi, “A novel machine learning method for estimating football players’ value in the transfer market”, *Soft Computing*, vol. 25, no. 3, pp. 2499–2511, Feb. 2021.
- [75] E. Arts. (Oct. 2016). “FIFA Video Games - Official EA Site”, [Online]. Available: <https://www.ea.com/en-gb/games/fifa> (visited on 2021-07-23).
- [76] B. Gerrard, “Analysing Sporting Efficiency Using Standardised Win Cost: Evidence from the FA Premier League, 1995 – 2007:” *International Journal of Sports Science & Coaching*, Mar. 2010.

- [77] P. Oskouie, S. Alipour, and A.-M. Eftekhari-Moghadam, “Multimodal feature extraction and fusion for semantic mining of soccer video: a survey”, *Artificial Intelligence Review*, vol. 42, no. 2, pp. 173–210, Aug. 2014.
- [78] N. Bhargava and F. Cuzzolin, “Challenges and opportunities for computer vision in real-life soccer analytics”, *CoRR*, vol. abs/2004.06180, 2020. arXiv: 2004.06180.
- [79] F. R. Goes, L. A. Meerhoff, M. J. O. Bueno, D. M. Rodrigues, F. A. Moura, M. S. Brink, M. T. Elferink-Gemser, A. J. Knobbe, S. A. Cunha, R. S. Torres, and K. A. P. M. Lemmink, “Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review”, *European Journal of Sport Science*, vol. 21, no. 4, pp. 481–496, Apr. 2021.
- [80] C. Cuevas, D. Quilón, and N. García, “Techniques and applications for soccer video analysis: A survey”, *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 29 685–29 721, Oct. 2020.
- [81] C. O. Fritz, P. E. Morris, and J. J. Richler, “Effect size estimates: current use, calculations, and interpretation.”, *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 2, 2012.
- [82] W. W. Cohen, R. E. Schapire, and Y. Singer, “Learning to order things”, *Journal of Artificial Intelligence Research*, vol. 10, pp. 243–270, 1999.
- [83] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine”, *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [84] A. Klementiev, D. Roth, and K. Small, “An unsupervised learning algorithm for rank aggregation”, in *European Conference on Machine Learning*, Springer, 2007, pp. 616–623.
- [85] A. Klementiev, D. Roth, and K. Small, “Unsupervised rank aggregation with distance-based models”, in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 472–479.
- [86] C.-G. Li, X. Mei, and B.-G. Hu, “Unsupervised ranking of multi-attribute objects based on principal curves”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 12, pp. 3404–3416, 2015.
- [87] C. R. Rao, X. Shi, and Y. Wu, “Approximation of the expected value of the harmonic mean and some applications”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 44, pp. 15 681–15 686, Nov. 2014.
- [88] P. M. E. Glickman. (2003). “Example of the Glicko-2 system”, [Online]. Available: <http://www.glicko.net/glicko/glicko2.pdf>.
- [89] V. Larivière and Y. Gingras, “Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation”, *Journal of Informetrics*, vol. 5, no. 3, pp. 392–399, Jul. 2011.

- [90] G. Google. (Jul. 2021). “Google Trends”. en-GB, [Online]. Available: <https://trends.google.com/trends/?geo=TR> (visited on 2021-07-13).
- [91] G. An, “The Effects of Adding Noise During Backpropagation Training on a Generalization Performance”, *Neural Computation*, vol. 8, no. 3, pp. 643–674, Apr. 1996.
- [92] G. F. Jenks, “The data model concept in statistical mapping”, *International Yearbook of Cartography*, vol. 7, pp. 186–190, 1967.
- [93] B. Torgler and S. L. Schmidt, “What shapes player performance in soccer? Empirical findings from a panel analysis”, *Applied Economics*, vol. 39, no. 18, pp. 2355–2369, Oct. 2007.
- [94] J. Cheng, Z. Wang, and G. Pollastri, “A neural network approach to ordinal regression”, in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, ISSN: 2161-4407, Jun. 2008, pp. 1279–1284.
- [95] M. W. Browne, “Cross-validation methods”, *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108–132, 2000.
- [96] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms”, *Neural Information Processing Systems*, vol. 25, 2012.
- [97] L. Breiman, “Random forests”, *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [98] J. H. Friedman, “Greedy function approximation: a gradient boosting machine”, *Annals of Statistics*, pp. 1189–1232, 2001.
- [99] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger, “Bagging survival trees”, *Statistics in Medicine*, vol. 23, no. 1, pp. 77–91, 2004.
- [100] G. Ridgeway, D. Madigan, and T. S. Richardson, “Boosting methodology for regression problems”, in *Seventh International Workshop on Artificial Intelligence and Statistics*, PMLR, 1999.
- [101] P. Smyth and D. Wolpert, “Linearly combining density estimators via stacking”, *Machine Learning*, vol. 36, no. 1, pp. 59–83, 1999.
- [102] A. N. Langville and C. D. Meyer, *Who’s #1?* Princeton University Press, 2012.
- [103] U. USCF. (1996). “About USCF”, [Online]. Available: <http://www.uschess.org/index.php/About-USCF/> (visited on 2020-11-12).
- [104] C. Spearman, “The Proof and Measurement of Association between Two Things”, *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, Publisher: University of Illinois Press.
- [105] M. G. Kendall, “A new measure of rank correlation”, *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

- [106] E. Yilmaz, J. A. Aslam, and S. Robertson, “A new rank correlation coefficient for information retrieval”, in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 587–594.
- [107] E. Yilmaz and J. A. Aslam, “Estimating average precision with incomplete and imperfect judgments”, in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 2006, pp. 102–111.
- [108] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques”, *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.
- [109] W. Safabun and K. Urbaniak, “A New Coefficient of Rankings Similarity in Decision-Making Problems”, in *Computational Science – ICCS 2020*, ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 632–645.
- [110] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: a highly efficient gradient boosting decision tree”, *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.
- [111] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: a next-generation hyperparameter optimization framework”, in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.
- [112] A. P. Sant’Anna, H. G. Costa, and V. Pereira, “CPP-TRI: a sorting method based on the probabilistic composition of preferences”, *International Journal of Information and Decision Sciences*, vol. 7, no. 3, pp. 193–212, 2015.
- [113] L. O. Gavião, A. P. Sant’Anna, G. B. A. Lima, and P. Garcia. (2018). “CPP: Composition of probabilistic preferences. R package version 0.1. 0”, [Online]. Available: <https://cran.r-project.org/package=CPP> (visited on 2020-05-03).
- [114] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
- [115] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, “A public data set of spatio-temporal match events in soccer competitions”, *Scientific Data*, vol. 6, no. 1, p. 236, Oct. 2019.
- [116] L. Wynants, B. V. Calster, G. S. Collins, and R. D. Riley, “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal”, *BMJ*, vol. 369, p. m1328, Apr. 2020, Publisher: British Medical Journal Publishing Group Section: Research.
- [117] H. Anton, “In-game strategy recommendations in association football: A study based on network theory”, M.S. thesis, Åbo Akademi University, 2019.



APPENDIX A

PLAYER ROLE DEFINITIONS

Aggressive CB

A central defender who is very aggressive at his approach. For them it is completely normal to make a high volume of fouls, specifically in the 33-66 area on the field. Due to their intensive gameplay, they tend to contribute with higher number of interceptions, aerial duels, clearances, and defensive duels. It is preferable for them to participate actively in the build-up of the attacks. Nowadays it is essential for any type of a CB to be able to play with the ball because most of the teams are relying on high press. Therefore, if the defender is not capable of playing on the ball, it immediately makes him a target for the opposition and a vulnerable piece of the team.

Stable CB

The stable CB has a more conservative approach. He relies mainly on his positioning. Speed is important, but it comes secondary to their awareness on the field and their ability to read the situation. There is a certain set of skills essential for CBs such as efficiency at winning aerial duels and defensive duels. For the modern-day football for almost all teams is it also essential for the defender to be able to contribute to the build-up of the attacks. There are leagues where predominantly it is not of such importance – for example, most of the Turkish clubs are searching for an aggressive CB and they do not have the appreciation of stable CBs.

Wingback

The wingback is a player most suitable in a formation with three CBs. Their main purpose is to provide depth in the final third of the opposition by creating chances via crosses and passes inside the penalty area. Their defensive contribution comes second, as it is expected for the CBs to cover the space left by the wingback. They have great stamina, durability. They are expected to have lower number of interceptions and lower success rate in defensive duels – around 50-55%. If they can play 1v1 (winning defensive duels) it is highly beneficial, but this is a secondary skill for them.

Full back The full back is a player who plays in formations with 4 defenders. Their main purpose is to provide coverage for the defensive line. Therefore, they are required to play well defensively and if they could provide width in the final third it is highly beneficial but not essential as it is a secondary skill.

Defensive midfielder

The defensive midfielder in most cases is the most aggressive player on the field. They are used as a shield for the defensive line when the team attacks. Nowadays we have seen a variety of different types of defensive midfielders – rock, deep-lying playmaker. The type of the player depends on the tactical approach by the coach. The rock is a very direct player, fast, with great ability to run tirelessly and cover a great portion of the field, sometimes even on his own when playing in a 4-1-4-1 formation. The rock is the player who needs to contribute most actively at winning the possession of the ball and to participate more in loose ball duels and defensive duels. Fouls matter to certain extent, but they can make a higher number than defenders, as they have much more interactions with opposition players.

Regista

The regista is a specific type of a defensive midfielder. He combines the responsibility of shielding the defense and build-up the attacks from deep. He collects the ball from the defenders and distributes it forwards in a quick fashion. The regista has excellent passing ability and long ball distribution. Furthermore, the regista could provide through balls as well for the speedy players who are waiting in a channel for a goal-scoring chance and exploit their explosiveness.

Holding midfielder

A midfielder who is responsible primarily for the build-up of the attacks. He is a player who mainly contributes defensively and in the build-up of the attacks. Their main skills are pace, durability, intensity, and defensive intelligence. They are well positioned in 1v1 situations and can intercept balls in key areas in midfield. It is not unorthodox for them to also provide coverage for the attacks from second position.

Box-to-box midfielder It is a midfielder who provides for the overall balance of the team in each element – defense, attack, and transition. They are very good at 1v1 duels, can intercept a good amount of balls which allows them to start counterattacking actions, and they could also participate actively in the finishing of the attacks, as they usually have a very good shooting ability. It is optional for them to participate in the build-up of the attacks.

Central Attacking midfielder

The attacking midfielder is a player mainly responsible for the development of the attacks in the final third of the opposition. They are the “brain” of the team and are the footballers who can create a high volume of goal-scoring opportunities and who can provide option for finishing. Due to the tactical development of the sport, nowadays the coaches have more demands for these players. In the past, they have had limited to no defensive duties, whereas now it is mandatory for them to track back players and to be able to play defensively 1v1.

Link-up striker

A striker who has a task not only to provide goals but also to get involved the other attacking players in the final third. They can play with the back towards the goal because of their physicality which makes them very difficult to guard by defenders. They actively participate in the build-up of the attacks in the final third and often create chances for the wingers and the other strikers.

Tank

Tall, physical striker who can win a high volume of aerial duels. They are particularly good at playing with the back towards the goal and to hold the ball. They are very good at finishing crosses and from set pieces. They don't have particularly good passing, and it is optional for them to link-up the play at the final third. Their main skills are related with their physicality and their ability to shoot. Pace is an optional skill as well.

False 9 It is a type of striker usually used in formation with three players up front, such as 3-4-3 or 4-3-3. This type of striker drops into the midfield and his main task is to create numerical advantage in the middle of the field. The false 9 usually is a striker with good shot, passing ability, pace and he can also provide goal-scoring chances for the other attacking players. They are good on the ball and could also play as an attacking midfielder.

Speedy striker

These types of strikers are usually spotted in Argentina. They are unorthodox players as they don't possess the typical physical features for a striker. They are short strikers, with great pace who are very aggressive and active at pressing opposition defenders. They have great shooting ability, can work fast on the ball, and can also provide goal-scoring chances for the other attacking players.

Second striker

Second striker is a hybrid position between a striker and an attacking midfielder. They provide support for the main striker for second position and they excel in tight spaces. The main striker is responsible mainly for the finishing at the penalty area and for creating space for the second striker to shoot. They usually are undersized and have explosiveness, speed, great shot and could contribute for the creation of goal-scoring chances.

Classic winger

The classic winger is a player who mainly relies on two key elements – dribbling and crossing. They are highly efficient at 1v1 duels against full backs and can dribble past them with ease. This allows them to exploit their ability to further cross and provide goal-scoring chances for the strikers inside the box. They could also finish the attacks, but it comes as a secondary skill. Pace is essential for this position and they could also track back oppositions. They are more of a supporting type of players in the attack.

Attacking winger

The main difference between a classic winger and an attacking winger is the fact that the attacking winger is a player with much more responsibility in finishing the attacks. They have great pace, shooting ability and can also provide goal-scoring chances for the other players up front. For them it is not of such a great importance to have good dribbling ability and crossing. Furthermore, it is not expected from them to track back oppositions, instead they are focused on high pressing.



APPENDIX B

FEATURE DEFINITIONS

Feature	Definition	Feature Family	
Avg Competition Rating	Avg rating of teams in player competition	Competition Properties	
Mean Player Value	Average player value of competition	Competition Properties	
Commonality 1	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 10	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 2	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 3	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 4	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 5	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 6	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 7	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 8	Injury commonality (1: Least, 10: Most)	Injuries	
Commonality 9	Injury commonality (1: Least, 10: Most)	Injuries	
Currently Injured	Whether the player is injured at the time of transfer	Injuries	
Severity 1	Injury severity (1: Least, 10: Most)	Injuries	
Severity 10	Injury severity (1: Least, 10: Most)	Injuries	
Severity 2	Injury severity (1: Least, 10: Most)	Injuries	
Severity 3	Injury severity (1: Least, 10: Most)	Injuries	
Severity 4	Injury severity (1: Least, 10: Most)	Injuries	
Severity 5	Injury severity (1: Least, 10: Most)	Injuries	
Severity 6	Injury severity (1: Least, 10: Most)	Injuries	
Severity 7	Injury severity (1: Least, 10: Most)	Injuries	
Severity 8	Injury severity (1: Least, 10: Most)	Injuries	
Severity 9	Injury severity (1: Least, 10: Most)	Injuries	
Domestic Cup Club 0 Alltime Player	All time player appearances in division 0 domestic cup	Player Appearances	
Domestic Cup Club 0 Alltime Player Difference	All time player appearances in division 0 domestic cup, compared to avg # in competition	Player Appearances	

Domestic Cup Club 0 Prevseason Player	Prev season player appearances in division 0 domestic cup	Player ances	Appear-
Domestic Cup Club 0 Pre- vseason Player Difference	Prev season player appearances in division 0 domestic cup, compared to avg # in competition	Player ances	Appear-
Domestic Cup Club 1 Alltime Player	All time player appearances in division 1 domestic cup	Player ances	Appear-
Domestic Cup Club 1 Alltime Player Difference	All time player appearances in division 1 domestic cup, compared to avg # in competition	Player ances	Appear-
Domestic Cup Club 1 Prevseason Player	Prev season player appearances in division 1 domestic cup	Player ances	Appear-
Domestic Cup Club 1 Pre- vseason Player Difference	Prev season player appearances in division 1 domestic cup, compared to avg # in competition	Player ances	Appear-
Domestic Cup Club 2 Alltime Player	All time player appearances in division 2 domestic cup	Player ances	Appear-
Domestic Cup Club 2 Alltime Player Difference	All time player appearances in division 2 domestic cup, compared to avg # in competition	Player ances	Appear-
Domestic Cup Club 2 Prevseason Player	Prev season player appearances in division 2 domestic cup	Player ances	Appear-
Domestic Cup Club 2 Pre- vseason Player Difference	Prev season player appearances in division 2 domestic cup, compared to avg # in competition	Player ances	Appear-
Domestic League Club 0 Alltime Player	All time player appearances in division 0 domestic league	Player ances	Appear-
Domestic League Club 0 Alltime Player Difference	All time player appearances in division 0 domestic league, compared to avg # in competition	Player ances	Appear-
Domestic League Club 0 Prevseason Player	Prev season player appearances in division 0 domestic league	Player ances	Appear-
Domestic League Club 0 Prevseason Player Difference	Prev season player appearances in division 0 domestic league, compared to avg # in competition	Player ances	Appear-

Domestic League Club 1 Alltime Player	All time player appearances in division 1 domestic league	Player appearances	Appearances
Domestic League Club 1 Alltime Player Difference	All time player appearances in division 1 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic League Club 1 Prevseason Player	Prev season player appearances in division 1 domestic league	Player appearances	Appearances
Domestic League Club 1 Prevseason Player Difference	Prev season player appearances in division 1 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic League Club 2 Alltime Player	All time player appearances in division 2 domestic league	Player appearances	Appearances
Domestic League Club 2 Alltime Player Difference	All time player appearances in division 2 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic League Club 2 Prevseason Player	Prev season player appearances in division 2 domestic league	Player appearances	Appearances
Domestic League Club 2 Prevseason Player Difference	Prev season player appearances in division 2 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic League Club 3 Alltime Player	All time player appearances in division 3 domestic league	Player appearances	Appearances
Domestic League Club 3 Alltime Player Difference	All time player appearances in division 3 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic League Club 3 Prevseason Player	Prev season player appearances in division 3 domestic league	Player appearances	Appearances
Domestic League Club 3 Prevseason Player Difference	Prev season player appearances in division 3 domestic league, compared to avg # in competition	Player appearances	Appearances
Domestic Super Cup Club 0 Alltime Player	All time player appearances in division 0 domestic super cup	Player appearances	Appearances
Domestic Super Cup Club 0 Alltime Player Difference	All time player appearances in division 0 domestic super cup, compared to avg # in competition	Player appearances	Appearances

Domestic Super Cup Club 0 Prevseason Player	Prev season player appearances in division 0 domestic super cup	Player ances	Appear-
Domestic Super Cup Club 0 Prevseason Player Difference	Prev season player appearances in division 0 domestic super cup, compared to avg # in competition	Player ances	Appear-
Domestic Super Cup Club 2 Alltime Player	All time player appearances in division 2 domestic super cup	Player ances	Appear-
Domestic Super Cup Club 2 Alltime Player Difference	All time player appearances in division 2 domestic super cup, compared to avg # in competition	Player ances	Appear-
Domestic Super Cup Club 2 Prevseason Player	Prev season player appearances in division 2 domestic super cup	Player ances	Appear-
Domestic Super Cup Club 2 Prevseason Player Difference	Prev season player appearances in division 2 domestic super cup, compared to avg # in competition	Player ances	Appear-
Domestic Super Cup Club 3 Alltime Player	All time player appearances in division 0 domestic super cup	Player ances	Appear-
Domestic Super Cup Club 3 Alltime Player Difference	All time player appearances in division 0 domestic super cup, compared to avg # in competition	Player ances	Appear-
Domestic Super Cup Club 3 Prevseason Player	Prev season player appearances in division 0 domestic super cup	Player ances	Appear-
Domestic Super Cup Club 3 Prevseason Player Difference	Prev season player appearances in division 0 domestic super cup, compared to avg # in competition	Player ances	Appear-
International Cup Club 0 Alltime Player	All time player appearances in division 0 international cup	Player ances	Appear-
International Cup Club 0 Alltime Player Difference	All time player appearances in division 0 international cup, compared to avg # in competition	Player ances	Appear-
International Cup Club 0 Prevseason Player	Prev season player appearances in division 0 international cup	Player ances	Appear-

International Cup Club 0 Prevseason Player Difference	Prev season player appearances in division 0 international cup, compared to avg # in competition	Player appearances	Appearances
International Cup Club 1 Alltime Player	All time player appearances in division 1 international cup	Player appearances	Appearances
International Cup Club 1 Alltime Player Difference	All time player appearances in division 1 international cup, compared to avg # in competition	Player appearances	Appearances
International Cup Club 1 Prevseason Player	Prev season player appearances in division 1 international cup	Player appearances	Appearances
International Cup Club 1 Prevseason Player Difference	Prev season player appearances in division 1 international cup, compared to avg # in competition	Player appearances	Appearances
International Cup Club 2 Alltime Player	All time player appearances in division 2 international cup	Player appearances	Appearances
International Cup Club 2 Alltime Player Difference	All time player appearances in division 2 international cup, compared to avg # in competition	Player appearances	Appearances
International Cup Club 2 Prevseason Player	Prev season player appearances in division 2 international cup	Player appearances	Appearances
International Cup Club 2 Prevseason Player Difference	Prev season player appearances in division 2 international cup, compared to avg # in competition	Player appearances	Appearances
International Cup International 0 Alltime Player	All time player appearances in division 0 international cup (national teams)	Player appearances	Appearances
International Cup International 0 Alltime Player Difference	All time player appearances in division 0 international cup, compared to avg # in competition (national teams)	Player appearances	Appearances
International Cup International 0 Prevseason Player	Prev season player appearances in division 0 international cup (national teams)	Player appearances	Appearances
International Cup International 0 Prevseason Player Difference	Prev season player appearances in division 0 international cup, compared to avg # in competition (national teams)	Player appearances	Appearances

International Cup International 1 Alltime Player	All time player appearances in division 1 international cup (national teams)	Player ances	Appear-
International Cup International 1 Alltime Player Difference	All time player appearances in division 1 international cup, compared to avg # in competition (national teams)	Player ances	Appear-
International Cup International 1 Prevseason Player	Prev season player appearances in division 1 international cup (national teams)	Player ances	Appear-
International Cup International 1 Prevseason Player Difference	Prev season player appearances in division 1 international cup, compared to avg # in competition (national teams)	Player ances	Appear-
International Cup International 2 Alltime Player	All time player appearances in division 2 international cup (national teams)	Player ances	Appear-
International Cup International 2 Alltime Player Difference	All time player appearances in division 2 international cup, compared to avg # in competition (national teams)	Player ances	Appear-
International Cup International 2 Prevseason Player	Prev season player appearances in division 2 international cup (national teams)	Player ances	Appear-
International Cup International 2 Prevseason Player Difference	Prev season player appearances in division 2 international cup, compared to avg # in competition (national teams)	Player ances	Appear-
International Su- per Cup Club 0 Alltime Player	All time player appearances in division 0 international super cup (national teams)	Player ances	Appear-
International Super Cup Club 0 Alltime Player Difference	All time player appearances in division 0 international super cup, compared to avg # in competition (national teams)	Player ances	Appear-
International Super Cup Club 0 Prevseason Player	Prev season player appearances in division 0 international super cup (national teams)	Player ances	Appear-
International Super Cup Club 0 Prevseason Player Difference	Prev season player appearances in division 0 international super cup, compared to avg # in competition (national teams)	Player ances	Appear-

International Super Cup Club 3 Alltime Player	All time player appearances in division 3 international super cup (national teams)	Player appearances	Appearances
International Super Cup Club 3 Alltime Player Difference	All time player appearances in division 3 international super cup, compared to avg # in competition (national teams)	Player appearances	Appearances
International Super Cup Club 3 Prevseason Player	Prev season player appearances in division 3 international super cup (national teams)	Player appearances	Appearances
International Super Cup Club 3 Prevseason Player Difference	Prev season player appearances in division 3 international super cup, compared to avg # in competition (national teams)	Player appearances	Appearances
Age At Time	Age at time of transfer	Player demographics	Demographics
Age Squared	Age^2 at time of transfer	Player demographics	Demographics
Birth Eu	Avg fee paid for transfer of the player up to date of transfer	Player demographics	Demographics
Birtharea Alpha3Code	Player birth area	Player demographics	Demographics
Foot	Foot of the player	Player demographics	Demographics
Height	Player height	Player demographics	Demographics
Passport Eu	Whether player has EU passport	Player demographics	Demographics
Passportarea Alpha3Code	Player passport area	Player demographics	Demographics
Avg Fee To Date	Avg fee paid for transfer of the player up to date of transfer	Player Transfers	Historical Transfers
Max Fee To Date	Max fee spent on the player by buying clubs to date of transfer	Player Transfers	Historical Transfers
Min Fee To Date	Min fee spent on the player by buying clubs to date of transfer	Player Transfers	Historical Transfers
Mv	Transfermarkt Market Value before transfer	Player Transfers	Historical Transfers
Num Paid Transfers To Date	Number of paid transfers the player has been part of to date	Player Transfers	Historical Transfers
Num Transfers To Date	Number of all transfers the player has been part of to date	Player Transfers	Historical Transfers
Season	Which year the transfer season is	Player Transfers	Historical Transfers
Tenure In Team	How long has the player been part of their team before transfer	Player Transfers	Historical Transfers

Total Fee To Date	Total fee expenditure on player's historical transfers	Player Historical Transfers	
Transfer Window	Winter/Summer transfer window	Player Historical Transfers	
Accelerations Reg	Acceleration regular variable from ranking API	Player Performance	Performance
Attacking Duels Str	Attacking duels str variable from ranking API	Player Performance	Performance
Crosses Str	Crosses str variable from ranking API	Player Performance	Performance
Days Since Last Match	Days since player's last match	Player Performance	Performance
Field Air Duels Str	Field air duels str variable from ranking API	Player Performance	Performance
Forward Passes Str	Fwd passes str variable from ranking API	Player Performance	Performance
Forward To Back And Vert Passes Rat	fwd passes ratio variable from ranking API	Player Performance	Performance
Fouls Reg	Fouls reg variable from ranking API	Player Performance	Performance
Headshots Reg	Headshots reg variable from ranking API	Player Performance	Performance
Interceptions Reg	Interceptions reg variable from ranking API	Player Performance	Performance
Key Passes Str	key passes str variable from ranking API	Player Performance	Performance
Linkup Plays Str	linkup plays str variable from ranking API	Player Performance	Performance
Long Passes Str	long passes str variable from ranking API	Player Performance	Performance
Looseballs Str	looseballs str variable from ranking API	Player Performance	Performance
Losses To Recoveries Rat	loss to rec ratio variable from ranking API	Player Performance	Performance
Net Expected Goals Reg	Net expected goals reg variable from ranking API	Player Performance	Performance
New Attacking Duels Str	new attacking duels str variable from ranking API	Player Performance	Performance
New Defensive Duels Str	new defensive duels str variable from ranking API	Player Performance	Performance
New Successful Dribbles Str	new successful dribbles str variable from ranking API	Player Performance	Performance
Opponent Half Recoveries Reg	Opponent half recoveries variable from ranking API	Player Performance	Performance
Own Half Losses Reg	Own half losses reg variable from ranking API	Player Performance	Performance

Passes Str	Passes str variable from ranking api	Player mance	Perfor-
Player Rank	Player rank from the ranking API (i.e. order of the rating)	Player mance	Perfor-
Player Rating	Player rating from the ranking API (actual averages)	Player mance	Perfor-
Pressing Duels Reg	Pressing duels regular variable from ranking API	Player mance	Perfor-
Progressive Runs Reg	Progressive runs regular variable from ranking API	Player mance	Perfor-
Received Passes Reg	Received passes regular variable from ranking API	Player mance	Perfor-
Recoveries To Fouls Rat	Recoveries to fouls variable from ranking API	Player mance	Perfor-
Shot Assists Reg	Shot assists regular variable from ranking API	Player mance	Perfor-
Shots Reg	Shots regular variable from ranking API	Player mance	Perfor-
Shots Str	Shots str variable from ranking API	Player mance	Perfor-
Successful Dribbles Str	Successful dribbles str variable from ranking API	Player mance	Perfor-
Through Passes Str	Through passes str variable from ranking API	Player mance	Perfor-
Xgassists Reg	xgassists reg variable from ranking API	Player mance	Perfor-
Xgshots Reg	xgshots reg variable from ranking API	Player mance	Perfor-
Player To Role Popularity Ratio	Ratio of player's historical search index to role's search index	Player Popularity	
Player Trend	Player's historical search index	Player Popularity	
Role Trend	Player's high-level role's historical search index	Player Popularity	
Agg Cb	Total % probability of agg_cb	Player Role	
Attacking Winger	Total % probability of attacking_winger	Player Role	
Cam	Total % probability of cam	Player Role	
Classic Winger	Total % probability of classic_winger	Player Role	
Code	Position code of the player	Player Role	
Fb	Total % probability of attacking_fb	Player Role	
Gk	Total % probability of gk	Player Role	
Holding Mid	Total % probability of holding_mid	Player Role	
Linkup Striker	Total % probability of linkup_striker	Player Role	
Regista	Total % probability of regista	Player Role	
Speedy Striker	Total % probability of speedy striker	Player Role	
Stable Cb	Total % probability of stable cb	Player Role	
Tank	Total % probability of tank	Player Role	
Unique Roles	Number of unique roles the player has played	Player Role	
Wb	Total % probability of wb	Player Role	

Domestic Cup Club 0 Alltime Team	All time team appearances in division 0 domestic cup	Team ances	Appear-
Domestic Cup Club 0 Prevseason Team	Prev season team appearances in division 0 domestic cup	Team ances	Appear-
Domestic Cup Club 1 Alltime Team	All time team appearances in division 1 domestic cup	Team ances	Appear-
Domestic Cup Club 1 Prevseason Team	Prev season team appearances in division 1 domestic cup	Team ances	Appear-
Domestic Cup Club 2 Alltime Team	All time team appearances in division 2 domestic cup	Team ances	Appear-
Domestic Cup Club 2 Prevseason Team	Prev season team appearances in division 2 domestic cup	Team ances	Appear-
Domestic League Club 0 Alltime Team	All time team appearances in division 0 domestic league	Team ances	Appear-
Domestic League Club 0 Prevseason Team	Prev season team appearances in division 0 domestic league	Team ances	Appear-
Domestic League Club 1 Alltime Team	All time team appearances in division 1 domestic league	Team ances	Appear-
Domestic League Club 1 Prevseason Team	Prev season team appearances in division 1 domestic league	Team ances	Appear-
Domestic League Club 2 Alltime Team	All time team appearances in division 2 domestic league	Team ances	Appear-
Domestic League Club 2 Prevseason Team	Prev season team appearances in division 2 domestic league	Team ances	Appear-
Domestic League Club 3 Alltime Team	All time team appearances in division 3 domestic league	Team ances	Appear-
Domestic League Club 3 Prevseason Team	Prev season team appearances in division 3 domestic league	Team ances	Appear-
Domestic Super Cup Club 0 Alltime Team	All time team appearances in division 0 domestic super cup	Team ances	Appear-

Domestic Super Cup Club 0 Prevseason Team	Prev season team appearances in division 0 domestic super cup	Team ances	Appear-
Domestic Super Cup Club 2 Alltime Team	All time team appearances in division 2 domestic super cup	Team ances	Appear-
Domestic Super Cup Club 2 Prevseason Team	Prev season team appearances in division 2 domestic super cup	Team ances	Appear-
Domestic Super Cup Club 3 Alltime Team	All time team appearances in division 0 domestic super cup	Team ances	Appear-
Domestic Super Cup Club 3 Prevseason Team	Prev season team appearances in division 0 domestic super cup	Team ances	Appear-
International Cup Club 0 Alltime Team	All time team appearances in division 0 international cup	Team ances	Appear-
International Cup Club 0 Prevseason Team	Prev season team appearances in division 0 international cup	Team ances	Appear-
International Cup Club 1 Alltime Team	All time team appearances in division 1 international cup	Team ances	Appear-
International Cup Club 1 Prevseason Team	Prev season team appearances in division 1 international cup	Team ances	Appear-
International Cup Club 2 Alltime Team	All time team appearances in division 2 international cup	Team ances	Appear-
International Cup Club 2 Prevseason Team	Prev season team appearances in division 2 international cup	Team ances	Appear-
International Cup International 0 Alltime Team	All time team appearances in division 0 international cup (national teams)	Team ances	Appear-
International Cup International 0 Prevseason Team	Prev season team appearances in division 0 international cup (national teams)	Team ances	Appear-
International Super Cup Club 0 Alltime Team	All time team appearances in division 0 international super cup (national teams)	Team ances	Appear-
International Super Cup Club 0 Prevseason Team	Prev season team appearances in division 0 international super cup (national teams)	Team ances	Appear-

International Super Cup Club 3 Alltime Team	All time team appearances in division 3 international super cup (national teams)	Team Appearances
International Super Cup Club 3 Prevsaison Team	Prev season team appearances in division 3 international super cup (national teams)	Team Appearances
Count Alltime Buying Free Transfer	# Of all time inbound free transfers of player's selling club	Team Historical Transfers
Count Alltime Buying Loan	# Of all time inbound loans of player's selling club	Team Historical Transfers
Count Alltime Buying Transfer	# Of all time inbound paid transfers of player's selling club	Team Historical Transfers
Count Alltime Selling Free Transfer	# Of all time outbound free transfers of player's selling club	Team Historical Transfers
Count Alltime Selling Loan	# Of all time outbound free transfers of player's selling club	Team Historical Transfers
Count Alltime Selling Transfer	# Of all time outpound free transfers of player's selling club	Team Historical Transfers
Count Prevsaison Buying Free Transfer	# Of prev season inbound free transfers of player's selling club	Team Historical Transfers
Count Prevsaison Buying Loan	# Of prev season inbound loans of player's selling club	Team Historical Transfers
Count Prevsaison Buying Transfer	# Of prev season inbound paid transfers of player's selling club	Team Historical Transfers
Count Prevsaison Selling Free Transfer	# Of prev season outbound free transfers of player's selling club	Team Historical Transfers
Count Prevsaison Selling Loan	# Of prev season outbound free transfers of player's selling club	Team Historical Transfers
Count Prevsaison Selling Transfer	# Of prev season outpound free transfers of player's selling club	Team Historical Transfers
Total Fee Alltime Buying Loan	All time total fee expenditure of player's club before transfer on loans	Team Historical Transfers
Total Fee Alltime Buying Transfer	All time total fee expenditure of player's club before transfer on transfers	Team Historical Transfers
Total Fee Alltime Selling Loan	All time total fee revenue of player's club before transfer on loans	Team Historical Transfers
Total Fee Alltime Selling Transfer	All time total fee revenue of player's club before transfer o transfers	Team Historical Transfers
Total Fee Prevsaison Buying Loan	Prev season total fee expenditure of player's club before transfer on loans	Team Historical Transfers

Total Fee Pre- vseason Buying Transfer	Prev season total fee expenditure of player's club before transfer on transfers	Team Historical Transfers
Total Fee Pre- vseason Selling Loan	Prev season total fee revenue of player's club before transfer on loans	Team Historical Transfers
Total Fee Pre- vseason Selling Transfer	Prev season total fee revenue of player's club before transfer o transfers	Team Historical Transfers
Team Comp Rat- ing Diff	Difference between rating of player's team be- fore transfer and avg competition rating	Team Properties
Team Rating	Rating of player's team before transfer	Team Properties





APPENDIX C

MISSINGNESS STATISTICS

Feature	Missing Ratio
max_fee_to_date	0.77562
min_fee_to_date	0.77562
avg_fee_to_date	0.77562
total_fee_to_date	0.77562
commonality_10	0.28875
severity_9	0.28875
severity_8	0.28875
severity_7	0.28875
severity_6	0.28875
severity_5	0.28875
severity_4	0.28875
severity_3	0.28875
commonality_1	0.28875
severity_10	0.28875
severity_1	0.28875
commonality_9	0.28875
commonality_8	0.28875
commonality_7	0.28875
commonality_6	0.28875
commonality_5	0.28875
commonality_4	0.28875
commonality_3	0.28875
commonality_2	0.28875
severity_2	0.28875
mv	0.07430