

Understanding Value in Association Football: A Computational Study on Player Valuation

Word Count: 10989

Abstract

In modern football, player valuation remains an important aspect, affecting both the economic strength and competitive standing of clubs. Although being such a crucial part of football, there are evident instances of clubs experiencing financial losses due to mispriced player acquisitions or sales. This research aims to understand in further detail the determinants of football player valuation, with a particular emphasis on understanding the importance of performance metrics, non-performance metrics, and popularity in determining a player's worth.

Three linear regressions were used to analyse these variables, producing specific coefficients for each. Additionally, a neural network was designed to use performance metrics to predict player positions, an important influencer of player valuation, addressing a current gap in the football industry where a standardised classification method is lacking. An adjusted market value (AMV) using results from the linear regressions as 'weightings' was developed to understand how large of an impact the different variables had on market value.

This study's findings found performance as the most important influencer on player valuation, with popularity and non-performance metrics being less significant. The neural network was able to predict a player's position accurately and this can be used to prove a player's position, and therefore influence their price. In alignment with existing literature, this study found a common theme where offensive players typically had higher valuations than their defensive counterparts, a pattern seen in the results where offensive positions and metrics had higher coefficients.

The results of this study have important implications in the football industry. Football agents and directors of football would be able to leverage such analytics to allow for accurate, and consistent valuation estimates that can be updated consistently.

Acknowledgements

A special thank you to Dr. Xiaolan Liu my dissertation supervisor, for her invaluable guidance, encouragement, and feedback throughout this research journey. Her expertise and insights have been instrumental in shaping this work.

Table of Contents

List of Figures.....	6
List of Tables.....	7
CHAPTER 1: INTRODUCTION.....	8
1.1 Background.....	8
1.2 Challenges In Player Valuation.....	9
1.3 Research Questions.....	10
1.4 Objectives.....	10
1.5 Research Significance.....	11
1.6 Research Limitations.....	11
CHAPTER 2: LITERATURE REVIEW.....	13
2.1 History of Football Player Valuation.....	13
2.2 Crowd-Sourced Market Valuation - TransferMarket.....	16
2.3 Position Classification.....	17
2.4 Reflective Summary and Future Directions.....	19
CHAPTER 3: RESEARCH METHODS.....	21
3.1 Research Questions & Conceptual Framework.....	21
3.2 Performance/ Non-Performance Based Analysis.....	22
3.2.1 Data Collection.....	23
3.2.2 Methodology.....	24
3.3 Player Popularity Analysis.....	25
3.3.1 Data Collection.....	25
3.3.2 Methodology.....	26
3.4 Position Classification Methodology.....	26
3.4.1 Data Collection.....	27
3.4.2 Methodology.....	27
3.5 Adjusted Market Value (AMV).....	28
CHAPTER 4: RESULTS.....	30
4.1 Performance & Non-Performance-Based Metrics Impact on Market Value.....	30
4.2 Player Popularity.....	35
4.3 Position Classification Using Neural Network.....	37
4.4 Adjusted Market Valuation (AMV).....	40
CHAPTER 5: DISCUSSION.....	42
5.1 Discussion of Results.....	42
CHAPTER 6: CONCLUSION.....	46
REFERENCES.....	48

APPENDIX A – WyScout Data 53
APPENDIX B – Further Results..... 56

List of Figures

Figure 1 - Concept of how TransferMarkt value players, based off Brunswik's Lens Model [27]. Adapted by Herm et al. [24]	17
Figure 2- Player salaries by position in the Bundesliga by Battre et al. [29].....	19
Figure 3- Conceptual framework mapping key indicators to player valuation.....	22
Figure 4 - Proposed methodology to identify impact of performance and non-performance metrics on player value.....	23
Figure 5 - Simple methodology to predict player positions from WyScout performance data.....	26
Figure 6 - Detailed architectural structure of the neural network used to classify player position based on performance metrics.....	28
Figure 7- Correlation between market value and collected metrics from WyScout.	33
Figure 8- Relationship between 'Goals' and Market Value	33
Figure 9 - Linear regression coefficients for position of a football player.....	34
Figure 10 - Market value of players for each position in the dataset.....	35
Figure 11 - Average number of search results by Position.....	36
Figure 12- Average Market Value by Popularity Bin.....	36
Figure 13- Relationship between Player Popularity and Market Value	37
Figure 14 - Optimised neural network architecture.	38
Figure 15 - Training loss and validation loss per epoch.....	39
Figure 16 - Relationship between Market Value and AMV.	40

List of Tables

Table 1 - Top 20 Premier League signings.....	10
Table 2 - Football player valuation studies and how many drivers of valuation [11] are analysed.	15
Table 3 - Sample of the pre-processed WyScout dataset.	30
Table 4 - Top 10 VIF values for performance metrics	31
Table 5- Top 10 VIF values for non-performance metrics	31
Table 6 - R-squared value of linear regression models for performance and non-performance-based metrics.....	31
Table 7 - Top 5 coefficients for performance-based linear regression	32
Table 8 - Bottom 5 coefficients for performance-based linear regression	32
Table 9 - Popularity Linear Regression Results	37
Table 10 - Top 10 trial results and their suggested hyperparameters.	38
Table 11 - Spot checks to see how network performs for players who are known to play different positions to their WyScout classified one.	40
Table 12- AMV and Market Value Analysis	41
Table 13- Regression coefficients for non-performance-based metrics.....	43
Table 14 - League value for top 5 European leagues. Source: TransferMarkt [47]	43
Table 15 - All regression models and their influence on market value.	44

CHAPTER 1: INTRODUCTION

Football player valuation involves a complex assessment of both quantitative and qualitative factors to determine a player's market value. It is a critical process in the world of football, incorporating many different types of variables such as age, skill, marketability and potential. In the era of data-driven decision making, the process of valuation has come to play a significant role in the strategic planning of clubs, influencing decisions on player transfers, contract negotiations and scouting strategies. It is vital for clubs to have an in-depth understanding and accurate application of player valuation methods to stay competitive, and more importantly, financially stable.

This chapter aims to introduce football player valuation through the discussion of the background and context, followed by an analysis of the current research problem, the research questions, research objectives, and finally the limitations of the study.

1.1 Background

Association Football, widely known as Football, holds the title as the world's most popular sport, boasting an estimated five billion fans across the globe [1]. The huge popularity and cultural importance of this sport has led to the development of a colossal economic industry, with the European Football Market alone valued at an impressive £25.1 billion [2]. This figure is expected to continue to increase due to the exponential growth of women's football in the last few years where research has shown the average viewing time per person in the UK has increased by more than double in the last year [6].

A noteworthy example of the escalating investments in football can be represented by Saudi Arabia's economic diversification plan, formally known as Saudi Vision 2030. As part of this strategy, multi-billion-pound investments are being funnelled into the football industry, further inflating the financial footprint of the sport [3].

Within this vast and dynamic market, determining the value of the players and their employment contracts – the sport's most important assets – is an essential task [11]. Accurate player valuation is crucial for football clubs, stakeholders, sponsors and even fans as it plays a vital role in transfer negotiations, contract negotiations and investment assessments.

1.2 Challenges In Player Valuation

Football player valuation is a complex area that currently poses a challenge in modern football. This can be seen in Table 1, which uses data sourced from TransferMarkt [4] to illustrate that out of the top 20 most expensive football signings in England's highest division, the Premier League, only one player was able to retain or improve the market value the player was purchased at, whereas majority of the other players had large drops in value.

It can clearly be seen that many football clubs are now valuing players at higher prices than ever before, however, the processes behind accurately valuing a player differ greatly between football clubs and data companies due to the large number of variables that impact price and how important each variable can be. As researched by Franceschi et al. [11], these variables can be performance-based metrics such as 'goals scored, passes completed, tackles made, position' or non-performance-based metrics such as 'age, height, nation'.

Traditional valuation methodologies, while effective to some extent, are often biased heavily towards readily observable metrics such as goals scored, or goals assisted. These methods tend to favour attacking positions, painting a simplistic and often misleading picture of a player's value. This is seen on Table 1, where only five of the most expensive Premier League transfers were defenders or goalkeepers. It can be argued that modern football exhibits greater fluidity and dynamism than ever before. This shift can be attributed to the transformative approaches of football manager Josep Guardiola. His philosophies, drawing significant inspiration from the Dutch concept of 'Total Football' championed by Johan Cruyff, have played a pivotal role in shaping the modern game [5]. Many teams now play free flowing football, where defensive players are expected to participate in offensive actions and, conversely, attacking players are required to contribute to defensive duties. This style of play blurs traditional role boundaries, raising questions about player valuation. Despite this shift, a discrepancy persists: a player classified as a defender who performs as offensively well as an attacker, will often have a lower market value. This disparity in market value, despite similar performance levels, calls for a deeper examination of the underlying factors in player valuation and how it could be better understood.

Table 1 - Top 20 Premier League signings

#	Player	Age (When Brought)	Season	Market Value (Before Transfer)	Current Market Value (Or Market Value If Sold)	Transfer Fee	Joined	Profit/Loss	Brought At A Price Lower Than Market Value?
1	Enzo Fernández	22	22/23	€ 55.00M	€ 80.00M	€ 121.00M	Chelsea FC	-41.00	No
2	Jack Grealish	25	21/22	€ 65.00M	€ 75.00M	€ 117.50M	Manchester City	-42.50	No
3	Declan Rice	24	23/24	€ 90.00M	€ 90.00M	€ 116.60M	Arsenal FC	-26.60	No
4	Romelu Lukaku	28	21/22	€ 100.00M	€ 40.00M	€ 113.00M	Chelsea FC	-73.00	No
5	Paul Pogba	23	16/17	€ 70.00M	€ 35.00M	€ 105.00M	Manchester United	-70.00	No
6	Antony	22	22/23	€ 35.00M	€ 60.00M	€ 95.00M	Manchester United	-35.00	No
7	Harry Maguire	26	19/20	€ 50.00M	€ 20.00M	€ 87.00M	Manchester United	-67.00	No
8	Jadon Sancho	21	21/22	€ 100.00M	€ 45.00M	€ 85.00M	Manchester United	-40.00	Yes
9	Romelu Lukaku	24	17/18	€ 50.00M	€ 75.00M	€ 84.70M	Manchester United	-9.70	No
10	Virgil van Dijk	26	17/18	€ 30.00M	€ 35.00M	€ 84.65M	Liverpool FC	-49.65	No
11	Wesley Fofana	21	22/23	€ 40.00M	€ 55.00M	€ 80.40M	Chelsea FC	-25.40	No
12	Darwin Núñez	23	22/23	€ 55.00M	€ 65.00M	€ 80.00M	Liverpool FC	-15.00	No
13	Kai Havertz	21	20/21	€ 81.00M	€ 55.00M	€ 80.00M	Chelsea FC	-25.00	Yes
14	Nicolas Pépé	24	19/20	€ 65.00M	€ 18.00M	€ 80.00M	Arsenal FC	-62.00	No
15	Kepa Arrizabalaga	23	18/19	€ 20.00M	€ 18.00M	€ 80.00M	Chelsea FC	-62.00	No
16	Kevin De Bruyne	24	15/16	€ 45.00M	€ 70.00M	€ 76.00M	Manchester City	-6.00	No
17	Kai Havertz	24	23/24	€ 55.00M	€ 55.00M	€ 75.00M	Arsenal FC	-20.00	No
18	Ángel Di María	26	14/15	€ 50.00M	€ 50.00M	€ 75.00M	Manchester United	-25.00	No
19	Rúben Dias	23	20/21	€ 35.00M	€ 80.00M	€ 71.60M	Manchester City	8.40	No
20	Casemiro	30	22/23	€ 40.00M	€ 40.00M	€ 70.65M	Manchester United	-30.65	No

1.3 Research Questions

Below are the questions this dissertation aims to answer:

1. How can a football player's position be classified based upon performance-based metrics?
2. How can performance and non-performance-based metrics impact a football player's valuation?
3. How can player popularity impact a football player's valuation?
4. How can these indicators be weighted in terms of importance to the value of a football player?

1.4 Objectives

With developments in artificial intelligence technologies and an increase in knowledge within big data, this dissertation aims to delve into the complexities of player valuation using statistical methods and neural networks to design an accurate model for player valuation.

The main objectives that this dissertation aims to achieve include:

1. To develop a neural network that uses performance-based metrics to classify a football player's position. The goal is to achieve an output that accurately reflects the player's role in a game based on their metrics and regardless of their classified starting position.

2. To analyse and quantify the relationship between both performance-based and non-performance-based metrics and a football player's valuation. This includes understanding which metrics (like goals scored, assists, age, nationality, height) significantly influence the market value of a player.
3. To quantify the popularity of a football player and analyse how it can have an impact on market valuation.
4. To create a valuation model that assigns appropriate weightings to performance-based metrics, non-performance-based metrics, position, and popularity indicators. The goal is to provide a comprehensive and fair understanding of a player's value that accounts for all these different variables.

1.5 Research Significance

This study hopes to contribute to the increasingly critical domain of sports economics. As mentioned by Zaytseva & Shaposhnikov [8], although both offensive and defensive contributions are equally important for a team to win a game of football, the football transfer market currently exhibits a sign of potential labour market inefficiency, where forwards have long enjoyed greater popularity and higher salaries compared to other players. This study will first address how to classify a player's position based on the player's data, and then how to better value a football player.

In a landscape where football is a multi-billion-pound industry, an empirically grounded understanding of player valuation is important for informed, strategic decision making among club managers, club owners, and football agents. It can lead to greater efficiency during transfer negotiations by providing evidence for a player's true valuation and can reduce overspending between clubs and even during player contract negotiations where a player may want a salary increase.

This dissertation can also provide benefits beyond the realm of football and other sports. By examining factors that shape player valuations, it can unveil the social and cultural elements that often play a part in these financial assessments, such as player nationality and player popularity.

1.6 Research Limitations

There are a few limitations to this study that need to be acknowledged for proper interpretation of the findings. Firstly, the consistency and methodology of football data collection poses significant challenges. A common method is using analysts which cover live matches and manually input when an action has been completed [9]. However, this is prone to human subjectivity and human error

where an analyst can believe that a ‘tackle’ is a ‘foul’ instead. This variation in data collection results in multiple different sources of football data, each with different accuracy levels and reliability. For this study, WyScout will be the primary source of data due to it also being used by many top clubs worldwide [10].

Secondly, the availability of accurate and complete football data is predominantly limited to the top-tier football leagues. It is very rare to get good coverage in lower leagues and in less popular countries. For this study, the data is collected from the top seven European football leagues. While this ensures a degree of data reliability, it can also introduce certain biases. The findings of this study could therefore be less likely to be applicable or accurate when extrapolated to lower leagues.

Finally, the football transfer market is a volatile environment. The datasets will use historical data from 2018 and this will not consider any external economic factors which may have caused changes in market values during those years, such as the Coronavirus disease (COVID-19) [7] and the Russo-Ukrainian War. The influence of the football club has on a player’s market value will also be out of scope for this study.

CHAPTER 2: LITERATURE REVIEW

The objective of this chapter is to critically review and synthesise existing literature relevant to the research questions, providing a theoretical foundation for this study. The focus of this research revolves around football player valuation and the various variables that can affect it. The literature review aims to explore four core areas: performance-based metrics, non-performance-based metrics, a player's classified position, and player popularity, with an emphasis on their influence in affecting a player's market valuation. Beyond this review's primary focus, there can be several other variables affecting player valuation that lie outside the scope of this review. This review will not cover aspects like football club's financial strategies or the impact of political events like Brexit on player valuation.

This literature review will commence with an investigation into the historical context and current methods of football player valuation, looking into both performance-based and non-performance-based factors. The dynamics of player popularity, social media presence and media portrayal in affecting a player's market value will be analysed, along with an exploration into traditional position classification and the potential of using neural networks in this area. This review will also critically assess the limitations of current valuation models, including challenges in machine learning applications, and will conclude with a look at future suggestions for how to value football players, and offer ideas on what might be explored or changed in this area in the future.

2.1 History of Football Player Valuation

Football player valuation has become a crucial part of the modern football industry, shaping transfer strategies and financial planning within clubs. Many research efforts have been undertaken to dig deeper into the variables that determine a player's value and to assess the significance of each contributing element. Francesci et al. [11] found that over the past three decades of football player valuation research, there were six primary drivers used in player valuation methods.

1. Time – Effects of time on player valuation. (Inflation)
2. Labour – Contract details of a player or any other valuations for a player.
3. Performance – How a player performs in matches. (Goals, assists, saves, etc.)
4. Club Characteristics – Club financial status, club sporting performance, club attendance.
5. Player Characteristics – Age, height, nationality, seniority etc.
6. Popularity – How popular a player is, social media following, media presence.

Research by Carmichael & Thomas in 1993 [12] found that historically, player valuation was mainly determined based on the club characteristics of the purchasing and selling clubs rather than a direct assessment of the player's characteristics and performance. The research suggested that the valuation process typically considered factors like club attendances, club league ranking and very straightforward performance metrics such as goals, which could enhance a selling club's position in negotiations. A study by Dobson et al. in 2000 [13], further reinforced the idea that player transfer fees were influenced by club characteristics but also by time effects and player characteristics. These studies had limitations due to the lack of availability of performance data during those times and did not state the importance of each variable towards the valuation of a player.

In 2005, with the availability of more football data, Tunaru et al. [14] set out to develop a theoretical financial methodology to answer the question of "How much is this player worth at this moment in time?" They did this using a purely performance-based analysis using many metrics that are based on a player's performance during games. The output was not a financial value, but an index that represent how good a player is, which could then provide an estimate for player valuation. This was the start of a shift in data-driven analysis for football player valuation as there were many more metrics available to be used, however the study was limited to the performance driver which only provided part of the answer to a player's true valuation.

In 2010, Frank and Nüesch [16] examined if talent or popularity can both contribute to a player's valuation. Using the concept of 'economics of superstars' by Rosen in 1981 [17], they set out to analyse how differences in talent or appeal can lead to differences in valuation. They discovered that both performance and popularity significantly contributed to market value differentiations in football. This comprehensive approach demonstrated for the first time that a player's valuation is not solely determined by performance or club characteristics, but also popularity. However, this study had its limitations, as it relied solely on data from the top German league, and the analysis was conducted based on a limited number of performance metrics. This narrow scope can restrict the generalisability of the findings and overlook and other factors that may affect player valuation.

With the study by Müller et al. [15] in 2017, a data-driven approach was used that integrated popularity, player characteristics and performance together to estimate player market values. The researchers found that using multilevel regression models could accurately estimate market values by comparing the output to actual transfer fees. This study was able to overcome limitations of older research by using a larger number of metrics for performance analysis and using social media data for popularity analysis. However, the study also had limitations such as bias as the dataset was primarily from the top five European leagues and the model was trained on TransferMarkt's estimates of market values.

In 2021, Poli et al. [18] presented an econometric approach to valuing football players. This research piece was slightly different to the ones above [12]-[17] as it prioritised the importance of the ‘labour’ and ‘player characteristics’ aspects that drive football player valuation. The study identified contract length, player experience, and age as the most important elements that affected transfer fees, and consequently, market values. The study was not without its limitations, the dataset used for analysing transfer fees among players can often be inaccurate or entirely unavailable, this is due to the private nature of these transfers. The econometric approach may not have been accurate for ‘superstar’ players who have their market values skewed by their popularities.

Whilst football player valuation research was primarily econometric as seen above, machine learning approaches have been emerging recently. Research conducted by Aydemir et al. in 2022 [19] used machine learning techniques to predict transfer values using predictive modelling. The model was able to accurately predict transfer fees rather than market values and proved it through comparisons with listed transfer fees on TransferMarkt. The research revealed that the model’s accuracy in predicting high profile transfer fees was inferior compared to its predictions for lower profile transfer fees. A limitation of the study was its failure to account for the ‘labour’ driver of valuation, which might have led to more accurate predictions of transfer fees for players with less than a year remaining on their contracts.

As seen by the research methodologies above [11]-[19] the way in which football players are being valued has evolved over time and data-driven approaches are now being used more often. As seen in Table 2, no research project examines all six drivers of valuation listed by Francesci et al. [11] due to the data being hard to obtain, but almost all research has a strong focus on player performance as a key indicator that drives value. A crucial element within player characteristics which studies have shown impact value is player position, however it is not understood how a player’s position is classified.

Table 2 - Football player valuation studies and how many drivers of valuation [11] are analysed.

Player Valuation Methodology	Time	Labour	Performance	Club Characteristics	Player Characteristics	Popularity
Carmichael & Thomas (1993) [12]	No	No	Yes	Yes	No	No

Dobson et al. (2000) [13]	Yes	No	No	Yes	Yes	No
Tunaru et al. (2005) [14]	No	No	Yes	No	No	No
Frank & Nüesch (2010) [16]	No	No	Yes	No	Yes	Yes
Müller et al. (2017) [15]	No	No	Yes	No	Yes	Yes
Poli et al. (2021) [18]	No	Yes	Yes	Yes	Yes	No
Aydemir et al. (2022) [19]	No	No	Yes	Yes	Yes	Yes
TransferMarkt [26]	No	Yes	Yes	Yes	Yes	Yes

2.2 Crowd-Sourced Market Valuation - TransferMarkt

As presented by Estellés-Arolas & González-Ladrón-de-Guevara [22], crowdsourcing is the practice of inviting a diverse group of individuals to voluntarily contribute skills or resources to complete a task. Both the requesting party and the participants benefit from the collaboration, and this could be financially or socially. The idea of crowdsourcing can be traced back to Galton's 1907 [23] observation which found that the collective decisions or predictions of a crowd might rival or even surpass those of an expert. This is more commonly known as 'wisdom of the crowd' or 'collective wisdom', and companies are now using this concept for football player market valuation.

TransferMarkt, a data company specialising in football, stands as a notable example of crowdsourced market valuation within the industry. They combine professional analysis with user-generated input, creating a comprehensive approach that leverages both expert analysis and the collective wisdom of enthusiasts. As mentioned by Herm et al. [24], TransferMarkt's crowdsourced market valuations often align closely with actual transfer fees, earning respect and credibility in the football industry. As seen in literature above, such valuations have become benchmarks for scientific research on football player valuation. The researchers also show how the concept of Brunswik's Lens Model [27] can be used to conceptualise how the TransferMarkt crowd predicts market value, as illustrated in Fig. 1. These crowdsourced valuations have also been used commonly in the media and this can influence the sports economy. The TransferMarkt valuations are also regularly used in real transactions and salary negotiations, further highlighting the role crowdsourcing plays in current football player valuation.

Research by Coates & Parshakov [25] reinforced the findings of [24], however stated that TransferMarkt market valuations can be biased predictors of true fee value and can present a misleading picture of a player’s true value.

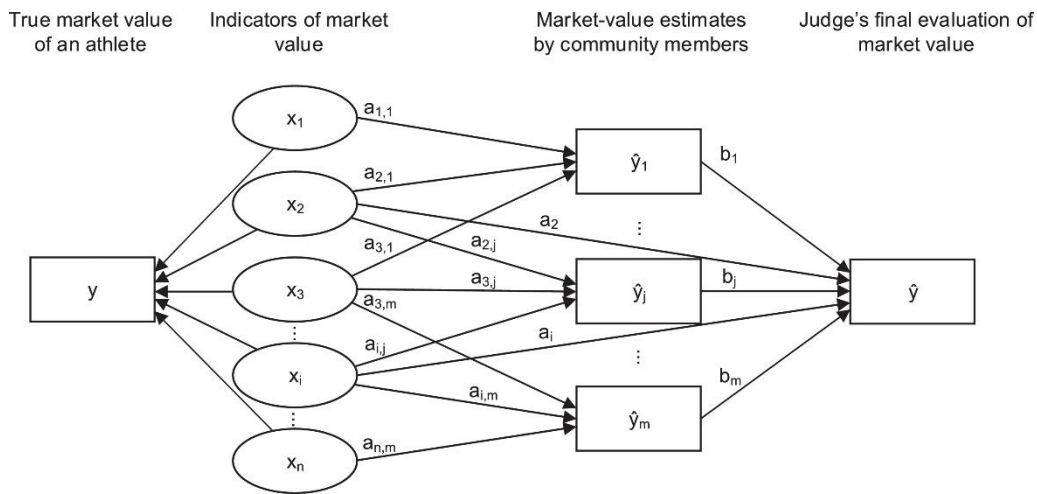


Figure 1 - Concept of how TransferMarkt value players, based off Brunswik's Lens Model [27]. Adapted by Herm et al. [24]

Compared to the econometric valuation methodologies seen in research above, TransferMarkt employ a unique methodology for player valuation, emphasising the community-driven approach rather than relying on algorithms. As disclosed by TransferMarkt administrators [26], the method integrates various pricing models and all key drivers of valuation listed by Francesci et al. [11] apart from ‘time’. TransferMarkt also incorporate situational conditions such as the financial pressure of a club, a player’s desires, or even actions such as a player going on strike in determining market valuation. These conditions can be purely speculative and do not fit neatly into quantitative data categories. Such variables aren’t commonly seen in existing literature, showing how the community’s subjectivity can be used to account for these situational conditions in player valuation.

While this methodology is reputable and commonly used in industry, it is not without limitations. The reliance on the wisdom of the community can lead to biased or inconsistent valuations, as seen in research by Coates & Parshakov [25]. The complexity of the factors involved can make it challenging for the crowd to arrive at a conclusion for a player’s market value, leading to potential inconsistencies or oversimplifications that may not reflect a player’s true value. Crowdsourcing can also lead to volatility, as public opinion can shift rapidly based on a player’s recent performances or media coverage. This can lead to fluctuations in a player’s price that may lead to inaccuracies.

2.3 Position Classification

The position of a player in football is considered under the branch of the ‘player characteristics’ driver of valuation, but it is a very complex area itself. From the studies outlined in Table 2 that involve

player position within their valuation methodologies, no study provides a consistent method detailing how position is classified and typically rely on the position classifications provided by their chosen dataset. Potential inaccuracies and bias in valuation methodologies can arise from these approaches, as demonstrated by He et al. in 2015 [28]. Their study highlighted that the responsibilities associated with each position differ, and given the goal-oriented nature of football, offensive performance metrics are often prioritised, possibly resulting in increased market valuations for attacking players.

In 2008, Battre et al. [29] conducted research on player salaries in the Bundesliga, Germany's top-tier football league. They found that forwards had the highest salaries, followed by midfielders, then defenders, with goalkeepers earning the least. This can be seen clearly in Fig. 2, which also shows an increasing player salary over time. This idea is reinforced by Deutscher & Büschemann in 2014 [30] who used concepts of game theory to explain that due to the 3-1-0-point system in football, which awards three points for a win, one point for a draw and none for a loss, encourages a more aggressive and riskier gameplay style, resulting in fewer draws and more victories. This can explain why offensive players are considered more valuable as their contributions can lead to higher points earned compared to defensive players or goalkeepers who are less likely to contribute to a win by scoring goals.

However, in 2022, Zaytseva & Shaposhnikov [8] disagreed with the notion that attacking actions were important to winning a game of football and explored the theory that defensive actions were underrated in comparison to offensive ones. They constructed econometric models to determine which actions held more significance in winning a game of football and concluded that both attack and defence were equally important to winning a game of football. They deduced that there could be a labour market inefficiency where offensive players are perceived as more valuable than defensive ones, even though both roles hold equal significance. However, for this study player field position classification was based on the datasets used for the study.

It is seen that position is an important variable within player valuation however existing literature primarily uses position classification already provided in datasets and do not question the nature of these classifications. This study aims to develop a methodology that can provide a position classification of higher accuracy using performance-based metrics. This is because of the fluid nature of modern football where we see that players are not often playing in their starting positions and often move around the pitch depending on if their team are in possession, out of possession or in transition.

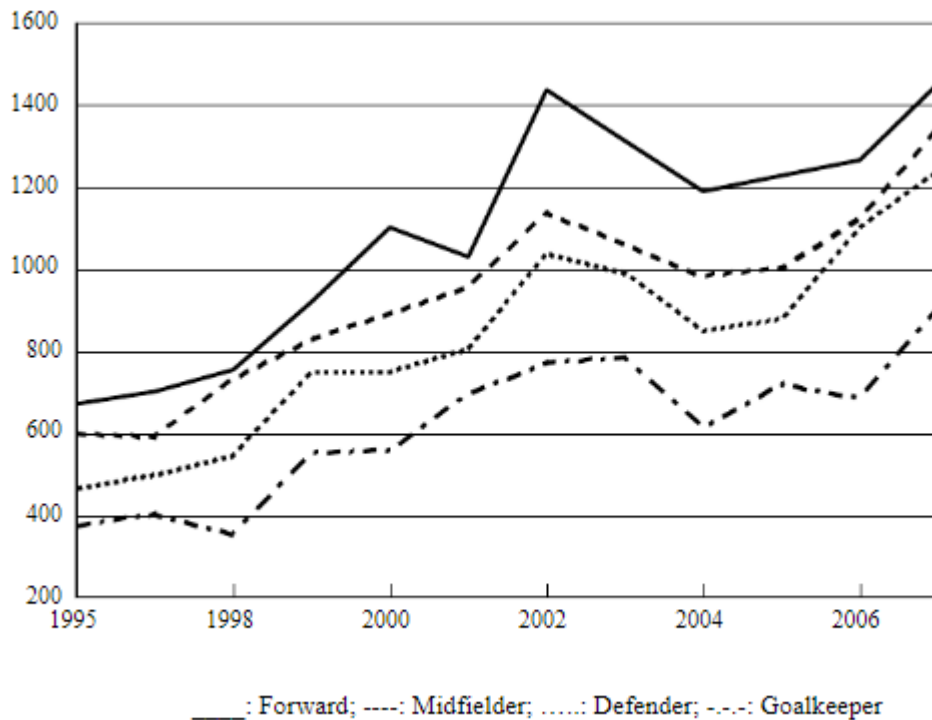


Figure 2- Player salaries by position in the Bundesliga by Battre et al. [29].

2.4 Reflective Summary and Future Directions

This literature review about football player valuation has provided a comprehensive understanding of the evolution and methodologies used in player valuation over time, from early econometric approaches to more recent data-driven and crowdsourcing methods. From the work by Francesci et al. [11], which identified six core drivers of valuation, to more recent contributions by Aydemir et al. [19] and Zaytseva & Shaposhnikov in 2022 [8], a clear transition to integrating multiple variables for player valuation is seen. Whilst performance remains a consistent indicator of player value, the introduction and recognition of ‘popularity’ by Frank and Nüesch [16] and ‘labour’ by Poli et al. [18] have increased the complexity of the issue. It remains to be seen which of these drivers are most important to valuation and a weighting system has not yet been researched.

The utilisation of crowdsourcing, as shown by TransferMarkt, represents the innovative blend of expert analysis and wisdom of the crowd. However, as critiqued by Coates & Parshakov [25], its reliance on community perspectives can lead to bias and volatility. Additionally, player position is seen as an important variable that impacts player valuation and salary. He et al. [28], showcases a potential of overvalue on offensive roles, potentially overshadowing the significance of defensive contributions, as argued by Zaytseva & Shaposhnikov [8].

Although there have been many advancements in player valuation methodologies, certain limitations still exist. The classification of a player's position has not yet been questioned and a consistent methodology for identifying it has not been seen. Many studies rely on pre-existing classifications in datasets without challenging the accuracy or nature as to how it has been derived, and there is clear evidence that these positions can impact player value as seen in research by Battre et al. [29] and Deutscher & Büschemann [30].

Given the dynamic nature of modern football, where player roles and positions are continually changing, there's an evident gap in the research of football player valuation that offers precise position classification based on performance metrics. Additionally, while current research methodologies explain the factors that drive valuation, the literature falls short in pinpointing which of these factors are most influential and how they should be proportionally weighted in player valuation. Addressing these gaps, the upcoming research aims to develop a position classification based on performance metrics. Additionally, it will determine how to properly weight player valuation drivers, specifically player performance, player popularity and player characteristics. This will provide a more comprehensive and clear approach to player valuation.

CHAPTER 3: RESEARCH METHODS

This chapter offers an in-depth exploration into the methods employed to investigate the complexities of position classification, performance, and popularity on football player valuation. The research questions, focused on player valuation and position classification, will be revisited. Through insights gained from the literature review, a conceptual framework tailored to this study will be presented. The subsequent sections will detail the research methodology, emphasising the use of neural networks and econometric analysis to evaluate player valuation. The chapter will wrap up by reviewing the limitations of the selected methodologies.

3.1 Research Questions & Conceptual Framework

As mentioned in Chapter 1, this study aims to answer the following research questions:

1. How can a football player's position be classified based upon performance-based metrics?
2. How can performance and non-performance-based metrics impact a football player's valuation?
3. How can player popularity impact a football player's valuation?
4. How can these indicators be weighted in terms of importance to the value of a football player?

Based on these questions, the proposed conceptual framework, depicted in Figure 3, emerges. The conceptual framework maps the network of factors that play a role in a football player's market value, mirroring the objectives of the study's research questions. Drawing from Wood et al. [31] foundational work on player roles based off on-field contributions, this framework highlights performance metrics as crucial for position classification, addressing Research Question 1 (RQ1). For Research Question 2 (RQ2), the relation of Performance and Non-Performance-Based Metrics with market value is presented. While earlier studies [12], [13], have documented the influence of player performance on market value, newer research suggests non-performance-based metrics like height or nationality can play a role in market valuation [15], [26].

Player popularity, a focus of Research Question 3 (RQ3) is included as an influential factor in this framework. Studies have shown that a player's media presence and brand, independent of their on-field performance, can command significant weight in the valuation process [15], [16], [19]. This shift, possibly due to the commercialisation of football [32], necessitates its inclusion in this player valuation framework.

The “Weight/Importance” element of this framework stems from Research Question 4 (RQ4). It suggests a weighting system where each factor has a unique and variable influence on a player’s market valuation. By capturing these weightings, this framework can signal the movement beyond simple interpretations of player valuation to a more complex understanding of it.

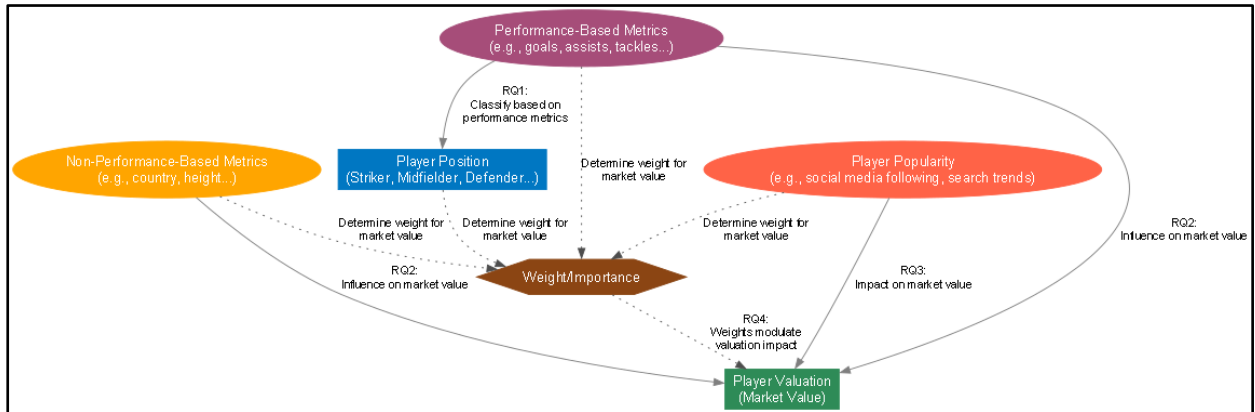


Figure 3- Conceptual framework mapping key indicators to player valuation.

3.2 Performance/ Non-Performance Based Analysis

In this study, performance and non-performance-based metrics are analysed to understand the relationship they have with market value. For a comprehensive analysis, data is first collected through the WyScout platform. Regression models are developed to capture the relationship among these metrics, offering insight into key determinants of a player’s market value and with the aim to outperform industry standards. Figure 4 shows the proposed methodology.

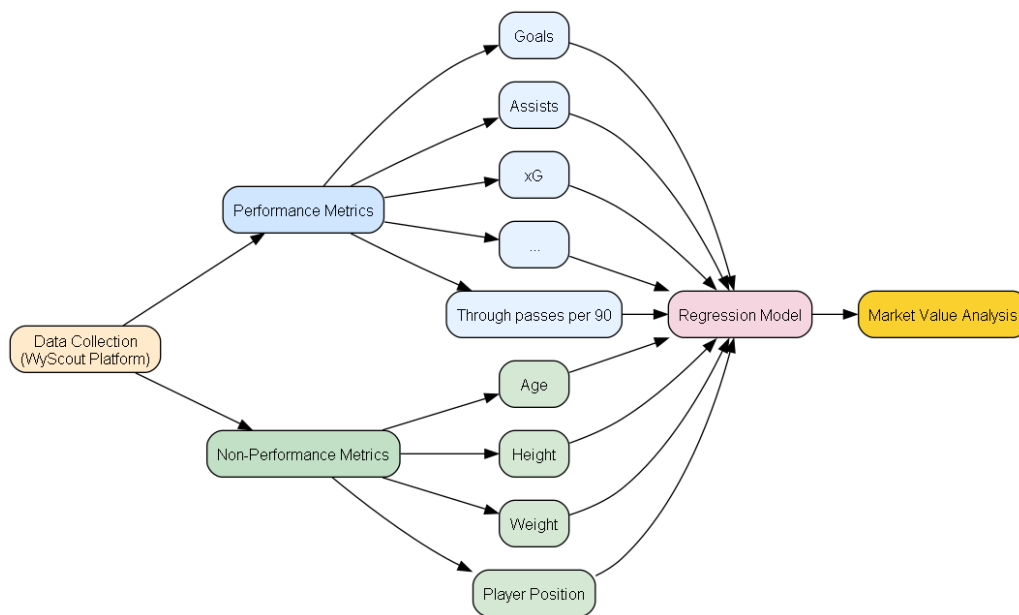


Figure 4 - Proposed methodology to identify impact of performance and non-performance metrics on player value.

3.2.1 Data Collection

For this study, data is sourced from WyScout, a reputable platform known for its large library of football data. This dataset provides in-depth insights into in-game events along with non-performance-based attributes like height and weight. The information is sourced from WyScout and is exported as Excel (.xlsx) files. A sample of the dataset is presented in Appendix A. This platform is chosen due to ease of use, depth of data and its reliability, evident from its endorsement by elite football clubs such as Manchester United and Arsenal [33]. To keep aligned with established research practices, such as the approach taken by Brandes & Franck [34], a minimum playtime condition is included to guarantee data consistency and reliability. For this study, players with a minimum of 900 minutes are selected, roughly equivalent to 10 complete matches. This ensures that the data sample largely represents regularly fielded players, therefore excluding those who might make sporadic appearances or those sidelined with injury. WyScout position data also tends to be complex, with multiple positions assigned to single players, for simplicity, WyScout positions are converted to simple positions based on a designed dictionary which can be found in Appendix A.

Data is also taken from the top seven European leagues [35], as these leagues tend to have higher data availability compared to their lower-tier counterparts. However, due to data being collected from only these top-tier leagues, the results might not be directly applicable to lower-tier or grassroots football scenarios.

In total 57 columns of data are collected for roughly 2000 players resulting in approximately 114,000 data points which are analysed. A detailed explanation of the metrics that are analysed can be found in Appendix A.

3.2.2 Methodology

There are many statistical methods available that can study the relationship between variables. Some methods such as simple linear regression can analyse the relationship between two variables. For this study, RQ2 asks “How can performance and non-performance-based metrics impact a football player’s valuation?” Given the type of data from WyScout and the need to consider multiple variables at once, multiple linear regression is chosen as the method, an approach also seen in [15].

Other methods were also investigated, while decision trees can handle multiple variables simultaneously, as demonstrated by Song & Lu [40], their added complexity makes them less fitting for this study. The researchers also found that strong correlation between variables can lead to inaccuracies, in the context of football, majority of the metrics are likely to be correlated, again showing why this approach is not useful for this context. Another potential technique explored was using Support Vector Machines (SVMs) which aim to find a hyperplane that best divides a dataset into classes. However, these are also unsuitable for this study due to poor interpretability and issues with multicollinearity.

Multiple linear regression is chosen due to its ability to provide clear, interpretable coefficients for each metric, providing a direct understanding of how specific performance and non-performance-based metrics can impact player valuation, as articulated by James et al. in their book [41]. However, there are also limitations to using this approach. The technique performs on certain assumptions, including linearity, homoscedasticity, and the absence of multicollinearity [42]. These assumptions, if violated can lead to misleading results. The presence of outliers can also influence the regression model negatively, as highlighted by Rousseeuw & Leroy [43]. Additionally, issues with multicollinearity can affect multiple linear regression analysis, however as seen in Shrestha’s 2020 study [36], use of Variance Inflation Factor can be effective in detecting and removing these issues successfully.

Despite the listed limitations, the decision to use multiple linear regression is informed and deliberate. The technique is widely used across many research fields and can handle multiple variables at once, making it ideal for this study. Ensuring the key assumptions of this model are met, it can provide a quantitative insight into exactly how each metric from the WyScout dataset can influence a player’s value.

3.3 Player Popularity Analysis

Player popularity is seen as a driver of player valuation according to various academic studies seen above [11], [15], [16], [19], [20]. This section outlines the approach this research employs to examine player popularity and its influence on a player's market value, explaining the reasons for selecting this methodology.

3.3.1 Data Collection

WyScout data from Section 3.2.1 consisting of 'player name' and 'team' information is being reused for this analysis.

A metric to quantify popularity is also needed. The study by Aydemir et al. [19] uses Google Trends [37] to measure how often a specific player is searched using the Google search engine. Although a reliable method, it is deemed unsuitable for this study. This is primarily because Google Trends normalises their data, presenting metrics that are comparative only to a specific player's popularity, rather than providing absolute figures.

A method to web scrape X (formally known as Twitter) was also investigated, however due to API changes in 2023, it was also unsuitable to be used due to web scraping limits being added.

The chosen approach is using the Google Custom Search API, a service offered by Google that allows users to programmatically perform searches on the entire web. Unlike traditional web scraping methods, the API offers structured results. Concatenating a player's name with their team and adding "footballer" to the search query, it is possible to obtain a count of search results, indirectly reflecting a player's online presence. For instance, a query might be "R. Lewandowski FC Barcelona footballer". This method is also used in research by [48] & [24].

This produces a file with approximately 2000 footballers, along with their count of search results through Google search. Further analysis can then be conducted as to how search results impact market value.

This approach does come with its challenges. Relying on the count of search results does not necessarily correlate directly with a player's popularity. While other methods were deemed unfit for this study, leading to this selection, it is acknowledged that this approach has drawbacks. For example, footballers sharing names or having common names can skew the search results count, introducing potential errors in this study.

3.3.2 Methodology

In this study, a linear regression model is employed to understand the relationship between search counts and the market value of footballers. The reasons for choosing this approach can be seen in Section 3.2.2, along with the advantages and disadvantages of using this approach.

Data binning, also known as ‘bucketing’, is also used for this study. This technique groups a set of data points into smaller intervals or ‘bins’ to simplify analysis and make patterns easier to spot. However, this approach also has challenges. Determining optimal bin size is difficult, large bins can oversimplify data, while small bins might not effectively reduce noise and lead to inaccuracies in the results.

These methods are chosen to identify the influence of popularity on a player’s market value due to ease of interpretability while still capturing the nuances of the complex dataset.

3.4 Position Classification Methodology

A key driver of player valuation is the position a football player is classified as. This is reinforced in the studies by [28], [29], [30], however research has not yet developed a methodology to classify position accurately. The primary position in which a player typically begins a football game is often referred to as their designated position, this can be inaccurate in modern football. This section describes a methodology to accurately classify a player’s position using a neural network. Figure 5 shows the proposed methodology.

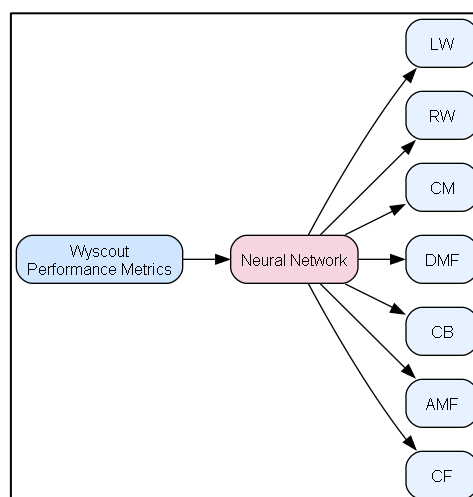


Figure 5 - Simple methodology to predict player positions from WyScout performance data.

3.4.1 Data Collection

For this analysis, the same WyScout dataset introduced in Section 3.1.1 is utilised. However, instead of focusing on the most recent 2022-2023 season, this dataset encompasses data spanning five seasons, beginning with the 2018-2019 season. This decision is chosen as to train a neural network effectively and enhance its accuracy, larger datasets are typically recommended. Leveraging data across multiple seasons not only provides a richer context, but also allows the network to recognise patterns over an extended period.

3.4.2 Methodology

In this study, RQ1 asks “How can a football player’s position be classified based upon performance-based metrics?”. To answer this question, a landscape of potential solutions is investigated, each accompanied with distinct advantages and challenges.

K-Nearest Neighbours (KNN) was investigated as a potential method to identify how to classify a player’s position based upon their performance metrics. The very simplicity that is KNN’s strength, is also one of its weaknesses. The model’s accuracy is closely tied to the chosen ‘k’, and as the volume of data increases, so does computational demand. This can make the model unstable and is more commonly known as the ‘curse of dimensionality’ which is seen to make KNN unstable as shown in research by Pestov in 2013 [44]. Due to the size of the dataset for this study, issues may arise if using KNN, therefore this method is deemed unsuitable for this study.

Logistic Regression, a statistical method viable for this classification problem, was also investigated. The method is simple to integrate and has easy interpretability, when applied to a football context, performance metrics such as passes, goals, and assists could be used as independent variables and their respective weights in the model would indicate their influence in predicting a player’s position. However, the linear nature of this model can limit its capability to capture more complex, nonlinear relationships that may be present in the dataset. It also assumes there is no multicollinearity, an assumption that is difficult to uphold due to the fluid nature of football, where most performance metrics will have some correlation between each other.

Among these methods, the use of neural networks is selected for this complex problem. These networks can distinguish complex relationships within large datasets, reducing reliance on manual feature engineering. As highlighted by Inan & Cavas [45], the use of neural networks can be very useful where there is an abundance of data, making it particularly useful for this study, which has a

collection of approximately 600,000 data points spanning across five seasons. However, the ‘black box’ nature of these neural networks can make interpretability harder, as it is not easy to answer ‘why’ a network has come to a certain solution.

Despite the limitations of neural networks, based on the examination of their capabilities compared to the other classification methods analysed along with insights from literature, neural networks are seen as suitable to use for this study to classify a player’s position based off their performance metrics. A detailed architectural structure of the proposed network can be seen in Figure 6.

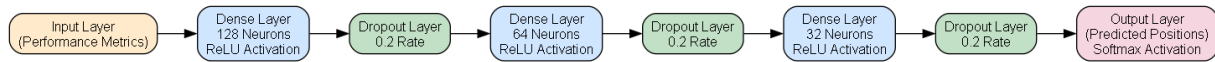


Figure 6 - Proposed architectural structure of the neural network to classify player position based on performance metrics.

3.5 Adjusted Market Value (AMV)

RQ4 asks, “How can these indicators be weighted in terms of importance to the value of a football player?” Through the linear regressions that are mentioned above, it is possible to generate an Adjusted Market Value (AMV) that can provide a nuanced valuation of a player, taking into consideration not only their current market value, but also their performance and non-performance-based metrics, their popularity, and their ‘predicted’ position.

The foundation for this adjustment is through the player’s initial market value. This is part of the WyScout dataset and is available to use, it can be represented as MV where MV denotes the Market Value. There are three major components of AMV:

Performance Impact: $PI = \sum_{i=1}^n (P_i \times C_i) * R_p^2$

- Where PI stands for the Performance Impact.
- P_i is each performance metric.
- C_i is the coefficient related to the i^{th} performance metric.
- R_p^2 is the R-squared value from the performance metric linear regression model.
- n represents the total number of performance metrics analysed.

Non-Performance Impact: $NPI = \sum_{j=1}^m (NP_j \times D_j) * R_N^2$

- Where NPI stands for the Non-Performance Impact.
- NP_j is each non-performance metric.
- D_j is the coefficient related to the j^{th} non-performance metric.
- R_N^2 is the R-squared value from the non-performance metric linear regression model.
- m represents the total number of non-performance metrics analysed.

Popularity Impact: $PoI = Po * E * R_{Po}^2$

- Where PoI stands for the Popularity Impact.
- Po is the popularity metric (search result count)
- E is the coefficient related to the popularity metric.
- R_{Po}^2 is the R-squared value from the popularity linear regression model.

Once these three metrics have been calculated, the Adjusted Market Value (AMV) can be simply formulated as:

$$AMV = MV + PI + NPI + PoI$$

This equation can provide a more comprehensive valuation of a player, capturing the elements of performance and non-performance-based metrics, the classified position of a player, and the popularity of a player.

There are some limitations to this approach, whilst AMV provides a data-driven approach to player valuation, it is still influenced baseline market value taken from WyScout, which can be seen to be inaccurate due to the limitations of position classification WyScout have. This formulation also uses R-squared values from the three linear regression models as adjustments, this can be inaccurate as a high R-squared value does not necessarily mean the model has a good fit, there could be cases of overfitting.

CHAPTER 4: RESULTS

In this chapter, quantitative findings on the variables that can affect football player valuation will be presented. The influence of both performance and non-performance-based metrics, as well as player popularity, on a player’s market value will be observed. Subsequently, the neural networks performance in determining a player’s position based off performance metrics will be analysed. The chapter will illustrate the findings of the AMV and will conclude by presenting a summary of the key findings most relevant to this study.

4.1 Performance & Non-Performance-Based Metrics Impact on Market Value

To answer RQ2, the WyScout dataset first had to undergo simple data preprocessing. Player position data was simplified for ease of analysis and was categorised into broader roles to allow for a more streamlined comparison of players across multiple leagues. A detailed table showing how the simplification was conducted can be found in Appendix A. The ‘standardscaler’ package from the scikit-learn library was also used to standardise features by removing the mean and scaling them to unit variance, this ensures that each feature has a mean of zero and standard deviation of one. This is beneficial for linear regression as it can help in mitigating the influence of features with larger scales, ensuring a consistent interpretation of coefficients.

The mathematical notation representing the operation of the ‘standardscaler’ is seen below:

$$x' = \frac{x - \mu}{\sigma}$$

where x' is the scaled value, μ is the mean, and σ is the standard deviation.

A sample of the pre-processed data can be seen in Table 3.

Table 3 - Sample of the pre-processed WyScout dataset.

	Player	Team	Position 1	Age	Market value	Birth country	Height	Weight	Goals	xG	Assists	xA	Duels per 90	Defensive duels per 90	Aerial duels per 90
0	Hugo Bueno	Wolverhampton Wanderers	LB	-1.650487	200000	Spain	-0.386910	-0.419228	-0.713102	-0.786288	-0.248585	0.123993	0.274465	1.140783	-0.420225
1	Jonny Otto	Wolverhampton Wanderers	RB	0.537946	17000000	Spain	-1.131609	-0.840882	-0.421577	-0.744785	-0.772397	-0.669157	0.500522	1.008275	-0.160069
2	C. Dawson	Wolverhampton Wanderers	CB	1.510582	2500000	England	0.804608	0.283530	-0.421577	-0.163734	-0.772397	-0.848628	-0.713768	-0.543347	0.752843
3	Matheus Nunes	Wolverhampton Wanderers	AMF	-0.677850	45000000	Brazil	0.208849	0.283530	-0.421577	0.078903	-0.248585	-0.061268	0.775455	0.503891	-0.254671
4	R. Ait Nouri	Wolverhampton Wanderers	LB	-1.407328	22000000	France	-0.386910	-0.840882	-0.421577	-0.435104	-0.772397	-0.466527	0.793784	1.752028	0.014945

Following data pre-processing, the features were categorised into two distinct groups: performance-based and non-performance-based metrics. VIF analysis was then conducted on these two groups respectively to determine the presence of multicollinearity among the factors. Due to the fluid nature of football, high VIF values were expected for numerous results such as ‘Expected Goals’ (xG) and ‘Goals’. A VIF value exceeding 10 is indicative of significant correlation between variables, Table 4 and Table 5 display the top 10 VIF values for performance metrics and non-performance metrics respectively. It is clearly seen that performance metrics have much higher correlated variables than non-performance metrics, the metrics with VIF values higher than 10 were then removed iteratively until no feature exceeded a VIF of 10. A detailed table of dropped columns and VIF values after removal can be seen in Appendix B.

Table 4 - Top 10 VIF values for performance metrics

	feature	VIF
44	Conceded goals	199851.899008
46	xG against	199451.671560
26	Passes per 90	46650.201625
30	Short / medium passes per 90	42306.299318
47	Prevented goals	4402.541022
17	Crosses per 90	2366.793016
31	Long passes per 90	1597.824422
19	Crosses from right flank per 90	1072.763630
18	Crosses from left flank per 90	1057.636462
8	PAdj Sliding tackles	94.679453

Table 5- Top 10 VIF values for non-performance metrics

	feature	VIF
1	Height	1.820600
2	Weight	1.793347
14	Position 1_GK	1.171592
3	Position 1_CB	1.103306
0	Age	1.053504
13	Position 1_AMF	1.046250
8	Position 1_LW	1.033289
11	Position 1_RW	1.030836
10	Position 1_RB	1.024167
4	Position 1_CF	1.020213

Two linear regressions were then trained on the two separate groups, aiming to understand the relationship between performance-based and non-performance-based metrics with market value. The dataset was split into training and testing sets, with an 80% and 20% split respectively, allowing the models to be tested on unseen data. This ensures the model can work with new data it has not yet seen, making sure it is not just good with the data it was trained with. Notably, the performance-based linear regression had an R-squared value of 0.341, indicating that it explained 34.1% of the variability in market value, whereas the non-performance-based regression had a lower R-squared of 0.053, suggesting that it accounted for just 5.3% of variability. This is seen in Table 6 below.

Table 6 - R-squared value of linear regression models for performance and non-performance-based metrics.

Regression Model	R-Squared
Performance-Based	0.341
Non-Performance-Based	0.053

The linear regression model focusing on performance metrics is detailed in Tables 7 and 8. Table 7 highlights the features with the top five coefficients, while Table 8 present the five lowest ones. Due to the WyScout data set being scaled, the coefficients represent the change in the market value for a one standard deviation increase in the predictor variable. An example of this, as seen on Table 7 can be for a player's 'lateral passes per 90' increasing by one standard deviation, their market value will rise £5,455,000. An increase of one standard deviation in performance metrics such as 'Goals', 'xG against', 'xA', and 'Key passes per 90' leads to the most significant rise in market value. Whereas the opposite holds true for performance metrics seen in Table 8.

Table 7 - Top 5 coefficients for performance-based linear regression

		coef	std err	t	P> t	[0.025	0.975]
20	Lateral passes per 90	5455000.0	5.79e+05	9.420	0.000	4.32e+06	6.59e+06
1	Goals	4965000.0	5.96e+05	8.328	0.000	3.8e+06	6.13e+06
29	xG against	2928000.0	5.97e+05	4.908	0.000	1.76e+06	4.1e+06
2	xA	2540000.0	6.9e+05	3.683	0.000	1.19e+06	3.89e+06
25	Key passes per 90	2390000.0	7.41e+05	3.226	0.001	9.37e+05	3.84e+06

Table 8 - Bottom 5 coefficients for performance-based linear regression

		coef	std err	t	P> t	[0.025	0.975]
23	Shot assists per 90	-4093000.0	8.36e+05	-4.893	0.000	-5.73e+06	-2.45e+06
12	Crosses from right flank per 90	-2269000.0	6.39e+05	-3.548	0.000	-3.52e+06	-1.01e+06
11	Crosses from left flank per 90	-2135000.0	6.15e+05	-3.471	0.001	-3.34e+06	-9.29e+05
21	Long passes per 90	-1523000.0	7.08e+05	-2.153	0.031	-2.91e+06	-1.35e+05
3	Defensive duels per 90	-1379000.0	5.31e+05	-2.595	0.010	-2.42e+06	-3.37e+05

Simple exploratory data analysis (EDA) was also conducted to understand the relationship between market value and the metrics collected from WyScout. This is clearly seen in Figure 7, where 'Age' is the most negatively correlated variable, and 'Goals' is the highest correlated variable. It is worth noting that these correlations are not particularly strong however, with the correlation for 'Age' and 'Goals' being -0.21 and 0.38 respectively. A detailed correlation matrix can be found in Appendix B.

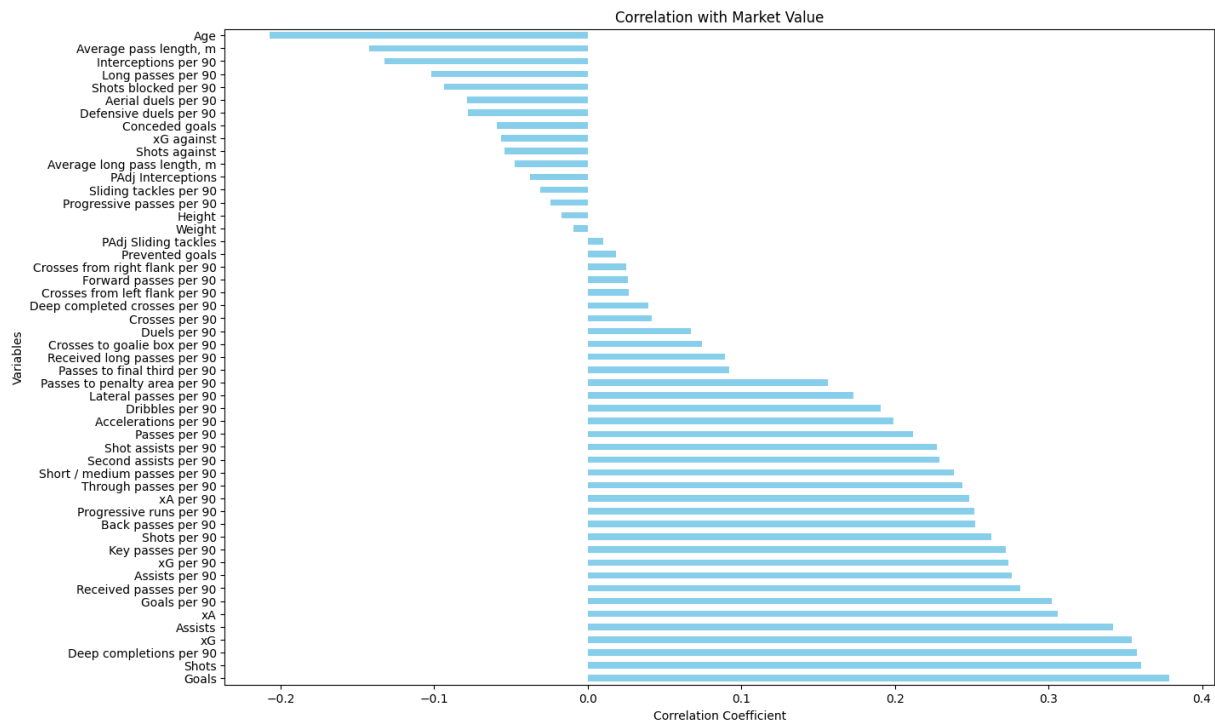


Figure 7- Correlation between market value and collected metrics from Wyscout.

Scatter graphs with the regression line for the metrics were also produced. Figure 8 shows the positive correlation between 'Goals' and 'Market Value' as well as the regression line indicating the general trend. It can be understood that an increase in 'Goals' clearly leads to an increase in player value. Further scatter graphs are seen in Appendix B.

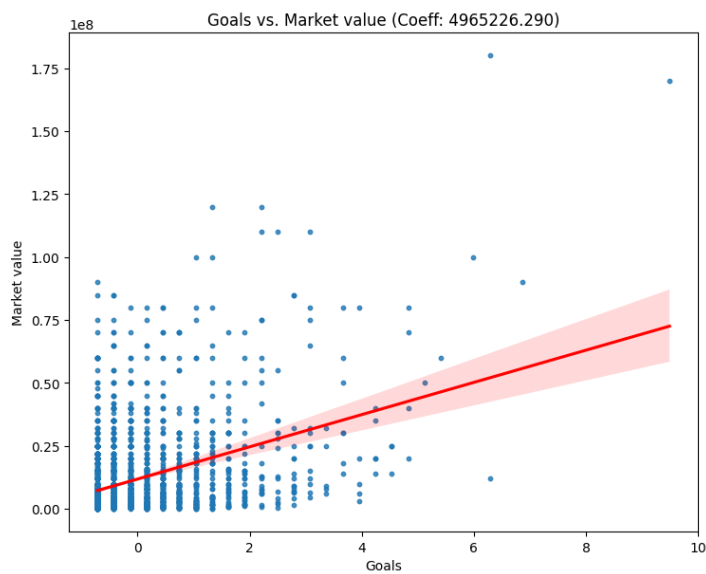


Figure 8- Relationship between 'Goals' and Market Value

The linear regression for non-performance metrics included position, Figure 9 shows the positions with the highest coefficients being ‘LW’, ‘CF’ and ‘DMF’ whereas the lowest being ‘RB’, ‘LWB’ and ‘GK’. Typically, attacking roles show a positive influence on market value, while defensive roles seem to have a negative effect. However, the high coefficient for ‘DMF’, a defensive role, is an unexpected observation.

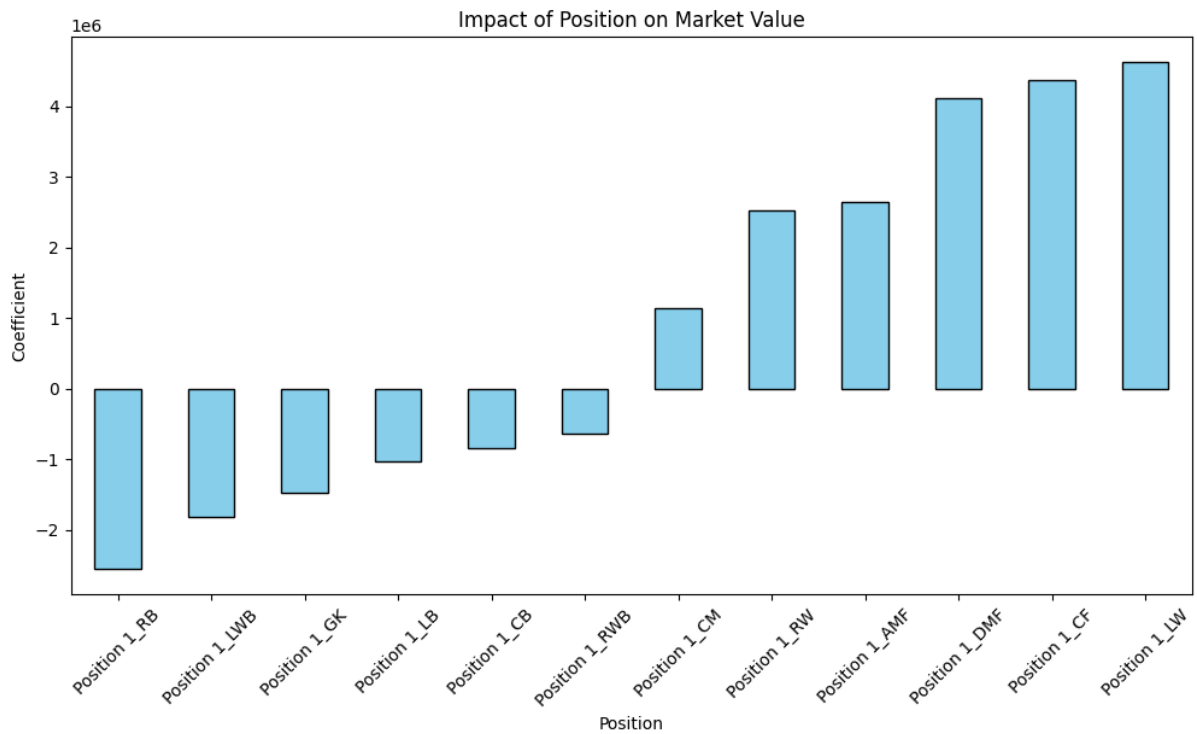


Figure 9 - Linear regression coefficients for position of a football player.

How much players are worth based on their ‘classified’ position was also analysed, as seen in Figure 10. It can clearly be seen the Centre Forward (CF) position has the highest value on average, whereas Right Back (RB) is the least valued position. The non-performance metric of ‘Birth country’ was also analysed as seen in Figure 11, it is seen that on average English players have the highest market value, followed by Argentinian and Croatian players respectively, this graph can be found in Appendix B.

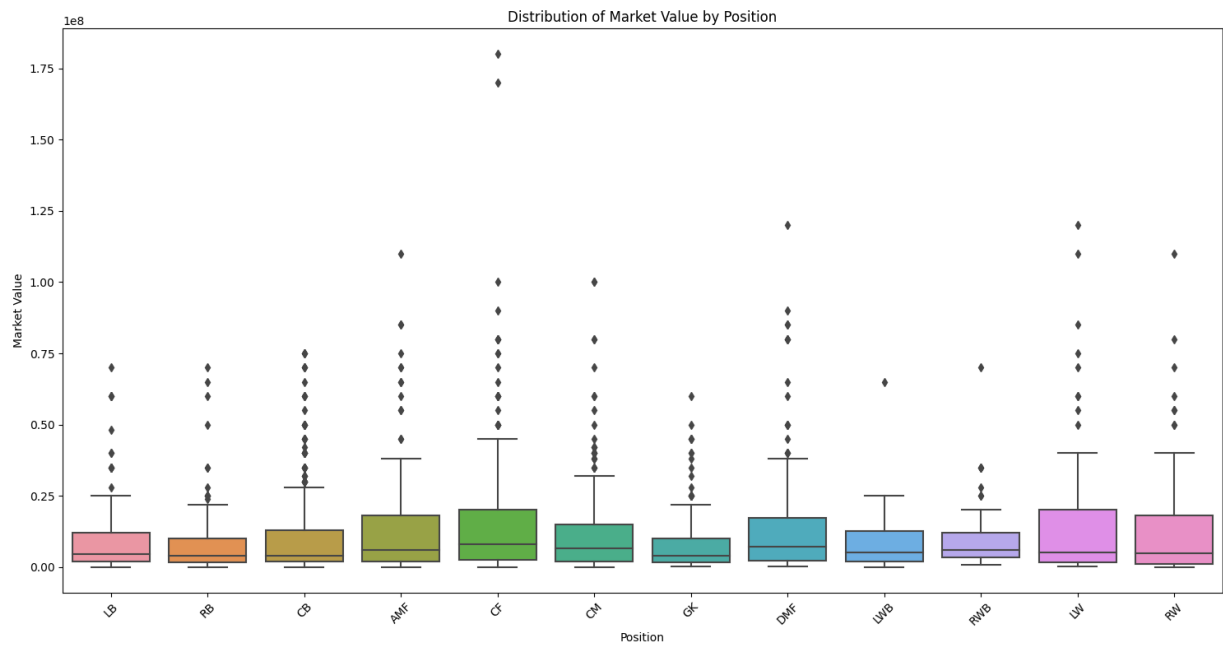


Figure 10 - Market value of players for each position in the dataset.

4.2 Player Popularity

As mentioned in Section 3.3.1, collecting data to indirectly reflect a player's popularity was challenging due to lack of available datasets and time constraints rendered manual data collection methods unsuitable for this study. Collecting the number of search results through Google was the chosen method and this was done by using the Google Cloud API and collecting search results data through a custom-made Google search engine. Detailed Python code to replicate this method, can be seen in Appendix B. Through this script, search result data was collected for every player in the data set. Figure 11 shows that Left Wing (LW) position has the players with the highest search count on average, with Right Wing-Back (RWB) being the lowest. As seen previously, offensive positions have higher search counts on average compared to defensive ones.

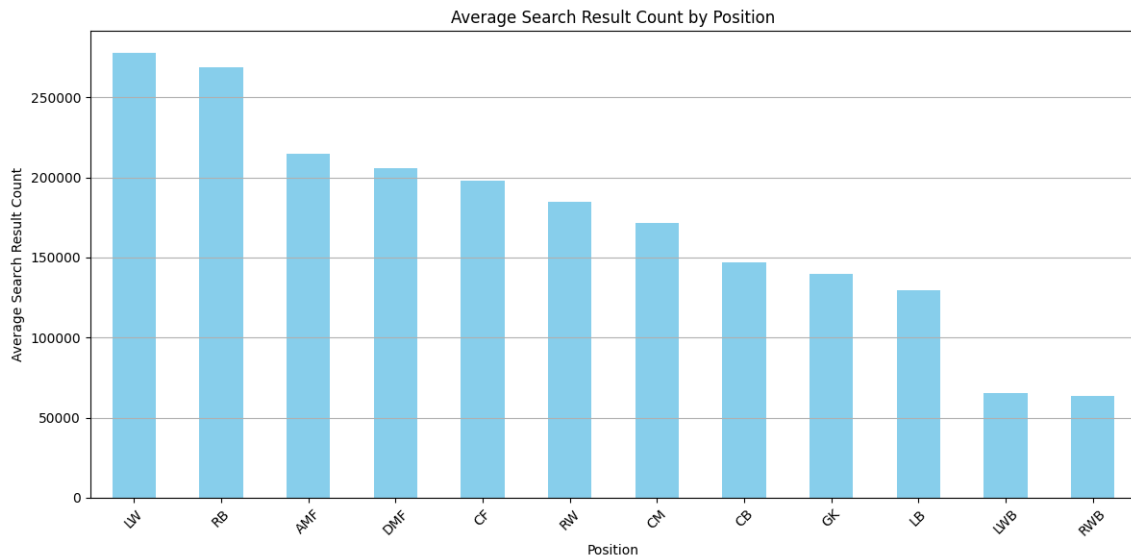


Figure 11 - Average number of search results by Position

The data was categorised into three distinct bins based on the distribution of ‘Search Results Count’, representing a player’s popularity. These bins were labelled as ‘Low’ (below the 33rd percentile), ‘High’ (above the 66th percentile) and ‘Medium’ (between the 33rd and 66th percentiles). As seen in Figure 12, for players within the ‘Low’ popularity bin, the average market value was approximately £4.24 million. Players in the ‘Medium’ popularity bin had an average market value of £9.96 million. Lastly, those in the ‘High’ popularity bin had an average market value of about £14.53 million.

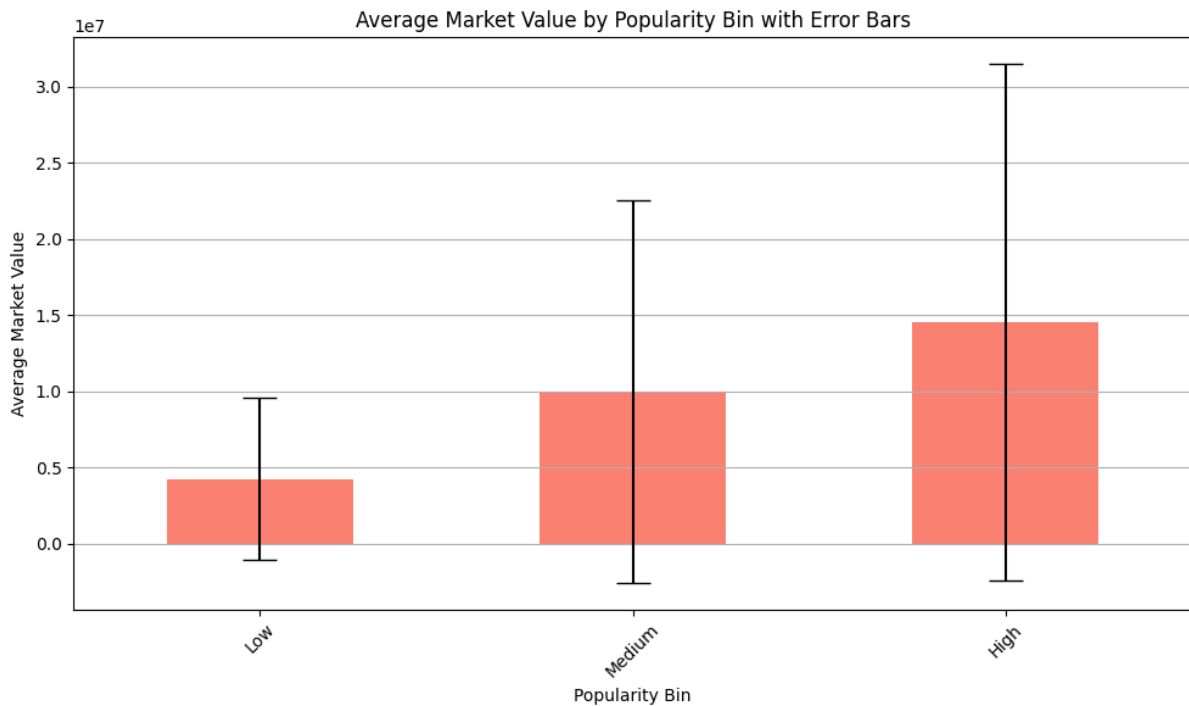


Figure 12- Average Market Value by Popularity Bin

A linear regression model was then fitted to the data with ‘Search results count’ being the independent variable, and ‘Market Value’ as the dependent variable.

Table 9 - Popularity Linear Regression Results

Linear Regression Model	R-Squared	Coefficient
Popularity	0.081	75.7573

This showed that the ‘search results count’ explained 8.1% of the variability in market value, and for each individual search result, market value would rise by £75.7573. A relationship can be seen in Figure 13.

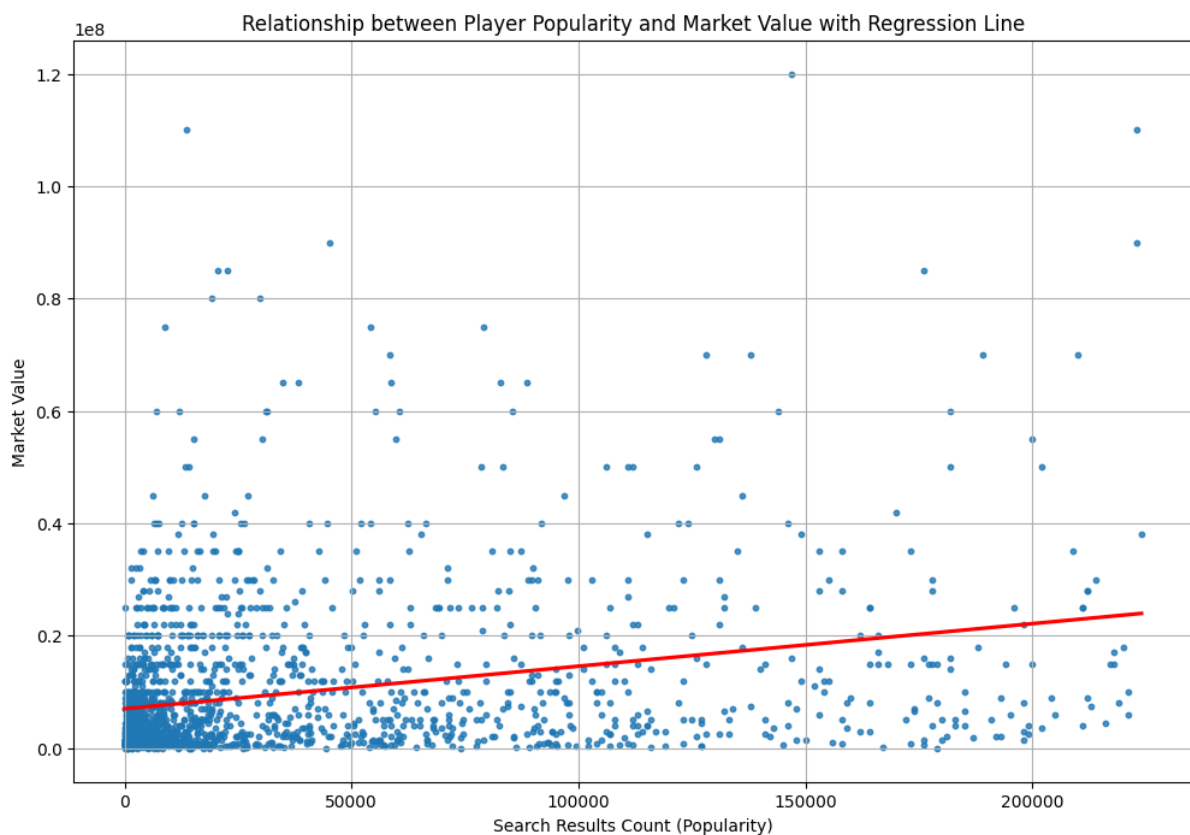


Figure 13- Relationship between Player Popularity and Market Value

4.3 Position Classification Using Neural Network

To train the neural network to classify positions based on performance metrics, five seasons worth of data from WyScout was used. The data was pre-processed and scaled in the same way as seen in Section 4.1. A sequential neural network was built using ‘TensorFlow’ [38]. The choice to use TensorFlow over PyTorch was influenced by the findings of Novac et al. [39], who found that

although PyTorch had faster training durations, TensorFlow demonstrated superior accuracy. The network was tested on the 22-23 season and trained on the four seasons before it.

The proposed architecture seen in Figure 6 was not used for this study. Hyperparameter optimisation techniques were used to find the most suitable configuration for this model. Using the ‘keras_tuner’ library, 254 trials were completed to find the most suitable parameters that gave the highest validation accuracy. Out of the 254 trials, the optimal configuration achieved validation accuracy of 80.3%. This configuration used a learning rate of 0.0005, an input layer with 96 neurons followed by a 15% drop out rate, and a hidden layer containing 128 neurons with a 30% dropout rate. The model architecture can be seen in Figure 14.

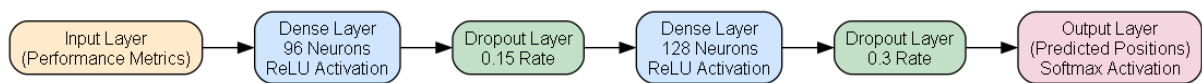


Figure 14 - Optimised neural network architecture.

Table 10 shows the suggested hyperparameters of the model for the top 10 highest performing trials, providing insights into the optimal configurations for this model.

Table 10 - Top 10 trial results and their suggested hyperparameters.

Trial No.	Units Input	Dropout 1	Num Layers	Units_0	Units_1	Units_2	Units_3	Dropout 2	Dropout 3	Learning Rate (2dp)	Score (5dp)
0208	96	0.15	1	128	192	160	96	0.30	0.40	0.0005	0.80307
0242	256	0.40	3	64	64	160	64	0.00	0.45	0.0006	0.80194
0250	64	0.30	4	128	64	128	96	0.20	0.15	0.0012	0.80137
0245	256	0.40	3	64	64	160	64	0.00	0.45	0.0006	0.80023
0207	96	0.15	1	128	192	160	96	0.30	0.40	0.0005	0.79966
0209	96	0.05	1	96	160	224	224	0.05	0.20	0.0007	0.79966
0253	224	0.45	1	64	224	192	128	0.25	0.30	0.0040	0.79852
0192	96	0.05	1	96	160	224	224	0.05	0.20	0.0007	0.79795
0203	96	0.05	1	96	160	224	224	0.05	0.20	0.0007	0.79795
0190	192	0.15	2	96	128	64	224	0.45	0.20	0.0008	0.79738

Training loss and validation loss was also calculated after 100 epochs as seen in Figure 15. The training loss was 0.3048 and the validation loss was higher at 0.6136. Although having a high validation loss compared to a training loss may seem as the model overfitting, this is expected due to

the model not expecting high accuracy as for football players who are playing out of position, the model is able to identify them even though their classified position is different. Out of 2149 players, the model predicted the same position for 1605 players, and predicted a different position for 544 players.

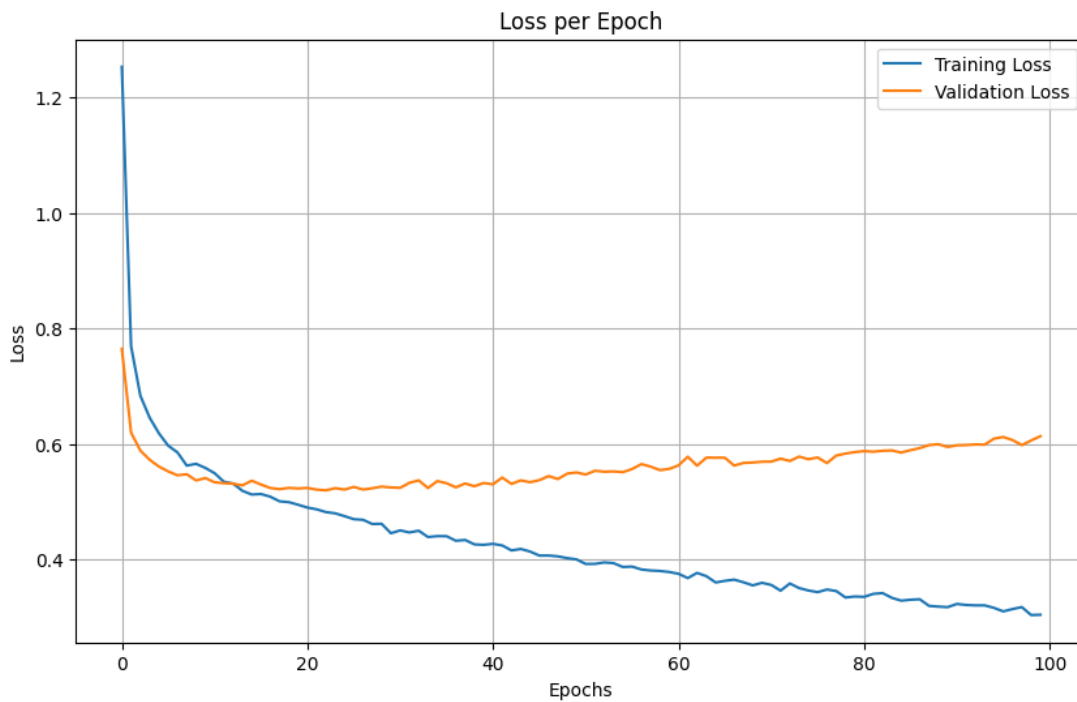


Figure 15 - Training loss and validation loss per epoch.

A confusion matrix was also created to visualise accuracy of the neural network in terms of classifying a player's position can be seen in Appendix B. As expected, the model performs with 100% accuracy when identifying Goalkeepers (GK), due to their metrics being unique such as 'Conceded goals'. A player's classified position and predicted position is usually matched apart from DMF where the network predicts those player's to be CMs instead. Spot checks were conducted to see how the network was performing for players who were known to play a different position compared to their WyScout classified position. Although not easy to prove without heatmap data, the model appears accurate in predicting positions for players as seen in Table 11.

Table 11 - Spot checks to see how network performs for players who are known to play different positions to their WyScout classified one.

Player	Team	WyScout Classified Position	Age	Market value	Predicted Position
Bernardo Silva	Manchester City	RW	28	80000000	CM
J. Milner	Liverpool	CM	37	2000000	RB
O. Zinchenko	Arsenal	LB	26	40000000	DMF
E. Camavinga	Real Madrid	LB	20	60000000	CM
Antony	Manchester United	AMF	23	70000000	RW
L. Messi	PSG	CF	35	50000000	RW

4.4 Adjusted Market Valuation (AMV)

The adjusted market value was then calculated using performance metrics, non-performance metrics, popularity, and the predicted positions along with all their coefficients calculated through the linear regressions. The AMV was compared to market value to see how it compared and can be seen in Figure 16.

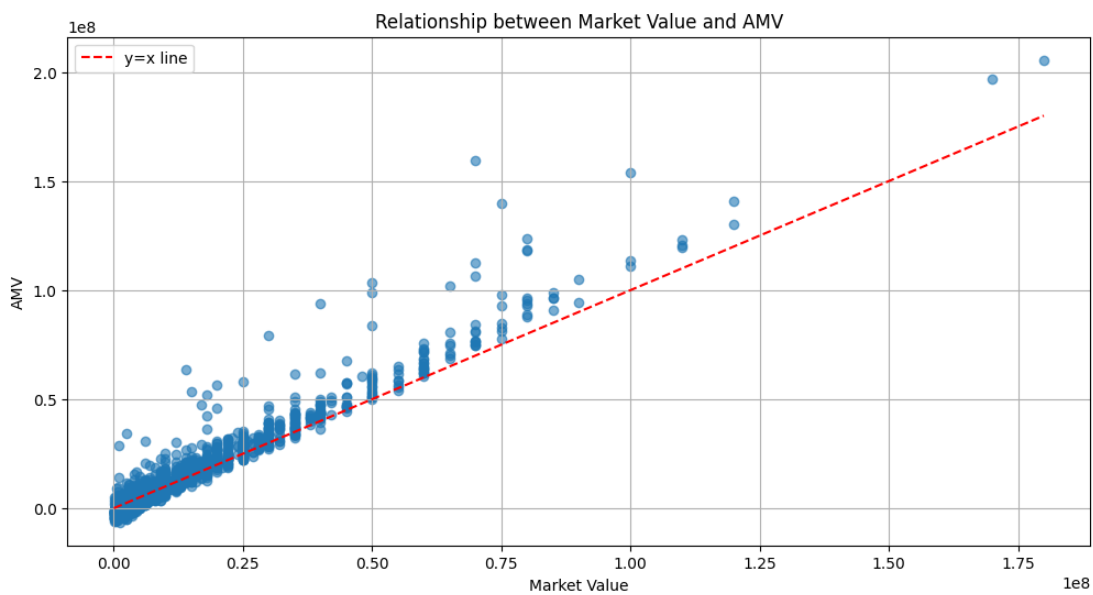


Figure 16 - Relationship between Market Value and AMV.

The market values were then split into quartiles and analysed as seen in Table 12. It is seen that AMV on average adds value to higher market value categories and reduces value of lower valued players.

Table 12- AMV and Market Value Analysis

Market Value Quartile	Average Market Value	Average AMV
Q1	-307,570.5	1,044,022
Q2	2,958,916	3,631,106
Q3	11,146,270	9,872,968
Q4	40,534,770	34,742,970

CHAPTER 5: DISCUSSION

The discussion chapter will interpret and explain the findings from the previous chapter relating back to the initial research questions proposed in Section 1.3. Findings will be compared to previous academic literature to understand how they fit into what is already known.

5.1 Discussion of Results

RQ1 asks “How can a football player’s position be classified based upon performance-based metrics?”. As seen in Chapter 4, a neural network was designed that used performance-based metrics to predict a player’s position. The optimised model was able to achieve an accuracy rate of 80.3% suggesting that the performance metrics chosen as input features have significant relevance to a player’s position on the field. The ‘inaccuracies’ of the model are not necessarily a bad thing, as seen in Table 11. The network was able to identify ‘true’ positions of footballers using their performance metrics in cases where their ‘classified’ position was either wrong, incorrect or had recently been changed.

Due to five seasons worth of data used by the neural network, the idea of trusting the network’s ‘inaccurate’ predictions can offer a unique approach to player valuation. If the model identifies a player as a striker based on performance metrics, but the player is officially a midfielder, it indicates that the player demonstrates qualities and skills typical for a striker. As seen in literature by Zaytseva & Shaposhnikov [8], offensive players are typically valued higher, an idea reinforced in research by [29] and [30]. This model can be used to prove that a player should be valued higher and could potentially be used by clubs and agents during negotiations.

A current limitation of this approach is that the dataset is based upon the top tiers of European football, meaning it may not be transferrable to lower league football, however this has not yet been tested. The neural network could also be continuously improved by increasing the amount of data available, as having more diverse training data can enhance the model’s predictive capabilities.

An interesting observation was that Defensive Midfielders (DMF) were often predicted to be Central Midfielders (CM) by the network, whereas for the rest of the position, there network was able to accurately predict most of them. There could be different reasons for this, but one key reason could be the similar role responsibilities. Both positions being centrally located and often involved offensively and defensively could make it hard for the network to distinguish between the two.

RQ2 asks “How can performance and non-performance-based metrics impact a football player’s valuation?” Two linear regression models were developed to answer this. Surprisingly, the

performance metric “Lateral passes per 90” emerged as having the most influence on market value. For every one standard deviation increase in this metric, the market value would rise approximately 5.45 million. This could be explained due to how modern football emphasises possession-based tactics, where retaining the ball and controlling tempo is crucial, players who execute these lateral passes accurately could be useful in such systems.

It is seen that in general, offensive metrics such as ‘Goals’, ‘expected assists’ (xA), ‘Key passes per 90’ have high coefficients, whereas the defensive metrics such as ‘Sliding tackles per 90’, ‘Defensive duels per 90’ and ‘Aerial duels per 90’ are usually on the lower side. This again, reinforces the idea that offensive players in general are valued higher than defensive players which is seen in literature above [28], [29], [30].

As seen in Table 13, the regression analysis for non-performance-based metrics found Age to negatively impact player value by around -4.6 million per standard deviation increase. This observation can make sense as footballing ability tends to decline as a footballer gets older, as found in a study by Rey et al. [46]. It is interesting to see a positive coefficient for ‘Weight’ and could be explained due to the physical nature of the game. The English Premier League (EPL) is generally seen as the most ‘physical’ league and is also the most expensive league in world football, as seen in Table 14 [47]. Players with higher weights are more likely to be in demand in such a league, where physical prowess can be an asset.

Table 13- Regression coefficients for non-performance-based metrics.

Non-performance Metrics	Coefficient
Age	-4,575,000
Height	-90,760
Weight	1,325,000

Table 14 - League value for top 5 European leagues. Source: TransferMarkt [47]









Competition	Country	Clubs	Player †	Avg. age †	Foreigners †	Forum	Total value †
First Tier							
 Premier League		20	577	26.4	67.9 %		€10.45bn
 LaLiga		20	501	27.4	41.3 %		€4.68bn
 Serie A		20	589	26.1	62.8 %		€4.67bn
 Bundesliga		18	527	25.6	48.8 %		€4.09bn
 Ligue 1		18	497	25.2	55.3 %		€3.48bn

Figure 9 also shows how position can affect market value and offensive positions are generally seen to have a higher influence on market value such as LW, CF, and AMF. Research from references [28],

[29], [30] further agrees with the higher valuation of offensive players. However, an interesting observation is the defensive midfielders (DMF) emerging as the third-highest coefficient in the analysis. There is not a clear answer as to why this is the case, however an explanation could be due to the scarcity of top tier DMFs, something seen in current football where defensive midfielders Declan Rice and Moisés Caicedo have been sold for over £100 million.

The linear regressions found that performance-metrics explained 34.1% of the variability in market value, whereas non-performance-based metrics explained 5.3% of the variability. This shows that performance metrics are indeed more important in player valuation, something that was expected prior to the research. However, non-performance metrics, although playing a lesser role, are still important in player valuation. A crucial non-performance aspect is the idea of ‘potential’ for a player, which can be seen as a large influencer of market value, however it cannot be quantified and the exclusion of ‘potential’ in this study can be seen as a huge limitation.

RQ3 analyses how player popularity can affect market value, the method used in studies [24] and [48] was replicated for this analysis using Google search counts. As seen in Figure 12, where the ‘binning’ technique was used, as search count increases, the average market value is seen to also increase. This agrees with research from previous studies that found player popularity to positively impact market value [15], [24], [48]. Figure 11 shows how position affects popularity, and as stated by Zaytseva & Shaposhnikov [8], offensive positions are generally seen to have higher search counts compared to defensive ones. An unexpected result is observed for the RB position where it seems to have the second highest popularity, but it was found due to the nature of some players in that position having common names and therefore inflating the search results count, proving limitations of using this methodology to indirectly quantify popularity.

Finally, RQ4 asks “How can these indicators be weighted in terms of importance to the value of a football player?” As seen on Table 15, the linear regressions found performance indicators to be the most important in impacting market value, followed by popularity and non-performance metrics.

Table 15 - All regression models and their influence on market value.

Regression Model	R-Squared	Variability In Market Value
Performance	0.341	34.1%
Non-Performance	0.053	5.3%
Popularity	0.081	8.1%

The Adjusted Market Value (AMV) metric was created to see how different factors within performance, non-performance, and popularity could change a football player’s market value. Based

on the results in Table 12, the AMV made the market values of less expensive players go down but improved the values of more expensive players. Certain players with lower market values were also seen with negative AMVs, which cannot be possible. It could also be argued that adjusting the AMV calculations to exclude WyScout market value as baseline, could prove to be more advantageous and provide a clearer image for a player's true worth.

CHAPTER 6: CONCLUSION

This study set out to investigate the different factors that can impact a football player's valuation. Four main questions were investigated in this study as seen in Section 1.3:

It was found that using a neural network and feeding five seasons worth of performance data, a neural network was able to identify a player's position to an accuracy of approximately 80%. However, for the context of this study, a 100% accuracy rate was not expected, and the 'inaccurate' predictions of the model did not necessarily mean the model was wrong.

Linear regressions were able to provide coefficients for each performance based and non-performance-based metric and it was seen that offensive metrics had higher coefficients than defensive ones on average. However, it is important to note that correlation does not imply causation. Offensive metrics having higher coefficients does not necessarily mean they directly improve market value, a player in an offensive position is more likely to have offensive metrics and higher popularity than a defensive player, leading to a skew in market value influence for offensive positions.

Non-performance metrics were also analysed, it was seen that 'age' negatively impacts market value and the position of a player also had influence depending on if it was an offensive position or a defensive position, this is where the neural network position classification could be used to prove a player's worth.

The study found player popularity to have a positive impact on market value, players who were in more offensive positions were generally more popular whereas the opposite held true for defensive players. Surprisingly, popularity was found to have a higher impact than non-performance indicators however this could be due to potential outliers in the dataset, especially for players with common names.

This research clearly shows that in terms of importance to market value, performance metrics rank highest, followed by popularity, with non-performance metrics being the least important. The AMV metric was developed to test the actual impact these indicators had on market value with the WyScout player value being used as a baseline.

There are also many limitations to this study. Firstly, this study uses data from the top seven European leagues where football is played at a high level. It may not be accurate for lower tier football, however gathering data for lower tier football can also be a challenge due to poor data availability [15]. There are also problems with multicollinearity due to football being a connected game. Almost every metric is related to another, an example can be player popularity can be correlated with position on a pitch which is then related to the performance metrics of a player. The data collected to quantify player

popularity is not as reliable as anticipated, a different method using social media followers should be explored for future research, however it would be difficult if using a large data set.

All six drivers of valuation mentioned by Franceschi et al. [11] are also not analysed in this study. There are key factors such as contract expiration date (labour), player potential, player injury history, club characteristics and the general economic landscape that have huge influence on player value but were out of scope for this study.

Football player valuation is an interesting area of research, especially in the current state that it is in with values rising at unprecedented rates. There is a lot more research that can be conducted in this area, especially in other areas that can drive valuation such as 'labour'. It can be argued that some of the most important drivers of valuation are unquantifiable such as 'potential' and 'personality', and research within this area can be of particular interest.

REFERENCES

[1]

FIFA, “The football landscape – The Vision 2020-2023 | FIFA Publications,” *FIFA Publications*. <https://publications.fifa.com/en/vision-report-2021/the-football-landscape/> (accessed Aug. 03, 2023).

[2]

L. Tantam, “European football market worth €28.4 billion (£25.1bn) as Premier League clubs lead the way to record revenues | Deloitte UK,” *Deloitte United Kingdom*, May 30, 2019. <https://www2.deloitte.com/uk/en/pages/press-releases/articles/european-football-market-worth-28-billion-euros-as-premier-league-clubs-lead-the-way-to-record-revenues.html> (accessed Aug. 03, 2023).

[3]

S. Kunti, “Saudi Soccer Bonanza: Public Investment Fund Backs Four Domestic Clubs To Grow Game And Influence,” *Forbes*, Jun. 06, 2023. <https://www.forbes.com/sites/samindrakunti/2023/06/06/saudi-soccer-bonanza-public-investment-fund-backs-four-domestic-clubs-to-grow-game-and-influence/?sh=6be847b534f4> (accessed Aug. 03, 2023).

[4]

TransferMarkt, “Premier League - Top Transfers,” *TransferMarkt*. https://www.transfermarkt.co.uk/premier-league/toptransfers/wettbewerb/GB1/plus//galerie/0?saison_id=alle&land_id=alle&ausrichtung=&spielerposition_id=alle&altersklasse=&w_s=&zuab=zu&art= (accessed Aug. 03, 2023).

[5]

A. Tweedale, “Coaches’ Voice | Pep Guardiola: In Others’ Words,” *The Coaches’ Voice*. <https://www.coachesvoice.com/cv/pep-guardiola-barcelona-torrent-manchester-city/> (accessed Aug. 03, 2023).

[6]

B. Grey, “Women’s sport: Research shows increase in viewers in 2022 - BBC Sport,” *BBC Sport*, Feb. 07, 2023. <https://www.bbc.co.uk/sport/football/64557964> (accessed Aug. 04, 2023).

[7]

R. Poli, L. Ravenel, and R. Besson, “The real impact of COVID on the football players’ transfer market,” Oct. 2020. <https://football-observatory.com/IMG/pdf/mr58en.pdf> (accessed Sep. 04, 2023).

[8]

I. Zaytseva and D. Shaposhnikov, “Moneyball in offensive versus defensive actions in football,” *Applied Economics*, no. 6, pp. 577–593, Jul. 2022, doi: 10.1080/00036846.2022.2091746.

[9]

AnalyiSport, “How Does The Premier League Collect Data? | AnalyiSport,” *AnalyiSport*, Sep. 01, 2022. <https://analyisport.com/insights/how-does-the-premier-league-collect-data/> (accessed Aug. 04, 2023).

[10]

N. Wright, “Wyscout: The scouting platform used by Arsenal, Manchester United and others | Football News | Sky Sports,” *Sky Sports*, Jan. 07, 2016. <https://www.skysports.com/football/news/11096/10120950/wyscout-the-scouting-tool-used-by-arsenal-manchester-united-and-others> (accessed Aug. 04, 2023).

[11]

M. Franceschi, J. Brocard, F. Follert, and J. Gouguet, “Determinants of football players’ valuation: A systematic review,” *Journal of Economic Surveys*, Feb. 2023, doi: 10.1111/joes.12552.

[12]

F. Carmichael and D. Thomas, “Bargaining in the transfer market: theory and evidence,” *Applied Economics*, no. 12, pp. 1467–1476, Dec. 1993, doi: 10.1080/00036849300000150.

[13]

S. Dobson, B. Gerrard, and S. Howe, “The determination of transfer fees in English nonleague football,” *Applied Economics*, no. 9, pp. 1145–1152, Jul. 2000, doi: 10.1080/000368400404281.

[14]

R. Tunaru, E. Clark, and H. Viney, “An option pricing framework for valuation of football players,” *Review of Financial Economics*, no. 3–4, pp. 281–295, Jan. 2005, doi: 10.1016/j.rfe.2004.11.002.

[15]

O. Müller, A. Simons, and M. Weinmann, “Beyond crowd judgments: Data-driven estimation of market value in association football,” *European Journal of Operational Research*, no. 2, pp. 611–624, Dec. 2017, doi: 10.1016/j.ejor.2017.05.005.

[16]

E. FRANCK and S. NÜESCH, “TALENT AND/OR POPULARITY: WHAT DOES IT TAKE TO BE A SUPERSTAR?,” *Economic Inquiry*, no. 1, pp. 202–216, Dec. 2010, doi: 10.1111/j.1465-7295.2010.00360.x.

[17]

S. Rosen, “The Economics of Superstars,” *The American Economic Review*, vol. 71, no. 5, pp. 845–858, doi: 10.2307/1803469.

[18]

R. Poli, R. Besson, and L. Ravenel, “Econometric Approach to Assessing the Transfer Fees and Values of Professional Football Players,” *Economies*, no. 1, p. 4, Dec. 2021, doi: 10.3390/economies10010004.

[19]

A. E. Aydemir, T. Taskaya Temizel, and A. Temizel, “A Machine Learning Ensembling Approach to Predicting Transfer Values,” *SN Computer Science*, no. 3, Mar. 2022, doi: 10.1007/s42979-022-01095-z.

[20]

“Professional Football Platform for Football Analysis - Wyscout,” *Wyscout*. <https://wyscout.com/> (accessed Aug. 09, 2023).

[21]

“TransferMarkt,” *TransferMarkt*. <https://www.transfermarkt.co.uk/> (accessed Aug. 09, 2023).

[22]

E. Estellés-Arolas and F. González-Ladrón-de-Guevara, “Towards an integrated crowdsourcing definition,” *Journal of Information Science*, no. 2, pp. 189–200, Mar. 2012, doi: 10.1177/0165551512437638.

[23]

F. GALTON, “Vox Populi,” *Nature*, no. 1949, pp. 450–451, Mar. 1907, doi: 10.1038/075450a0.

[24]

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*, no. 4, pp. 484–492, Oct. 2014, doi: 10.1016/j.smr.2013.12.006.

[25]

D. Coates and P. Parshakov, “The wisdom of crowds and transfer market values,” *European Journal of Operational Research*, no. 2, pp. 523–534, Sep. 2022, doi: 10.1016/j.ejor.2021.10.046.

[26]

TransferMarkt, “Transfermarkt Market Value explained - How is it determined? | Transfermarkt,” *Football transfers, rumours, market values and news | Transfermarkt*, May 13, 2021. <https://www.transfermarkt.co.in/transfermarkt-market-value-explained-how-is-it-determined-/view/news/385100> (accessed Aug. 10, 2023).

[27]

E. Brunswik, *Conceptual Framework Of Psychology*. University Of Chicago Press, 1952.

[28]

M. He, R. Cachucho, and A. Knobbe, “Football Player’s Performance and Market Value,” Jun. 2015. https://dtai.cs.kuleuven.be/events/MLSA15/papers/mlsa15_submission_8.pdf (accessed Aug. 10, 2023).

[29]

M. Battre, C. Deutscher, and B. Frick, “Salary Determination in the German ‘Bundesliga’: ,” *ResearchGate*, Jul. 2008. https://www.researchgate.net/publication/24131439_Salary_Determination_in_the_German_Bundesliga_A_Panel_Study.

[30]

C. Deutscher and A. Büschemann, “Does Performance Consistency Pay Off Financially for Players? Evidence From the Bundesliga,” *Journal of Sports Economics*, no. 1, pp. 27–43, Feb. 2014, doi: 10.1177/1527002514521428.

[31]

C. T. Woods, J. Veale, J. Fransen, S. Robertson, and N. F. Collier, “Classification of playing position in elite junior Australian football using technical skill indicators,” *Journal of Sports Sciences*, no. 1, pp. 97–103, Jan. 2017, doi: 10.1080/02640414.2017.1282621.

[32]

P. Kennedy and D. Kennedy, “Football supporters and the commercialisation of football: comparative responses across Europe,” *Soccer & Society*, no. 3, pp. 327–340, Mar. 2012, doi: 10.1080/14660970.2012.655503.

[33]

N. Wright, “Wyscout: The scouting platform used by Arsenal, Manchester United and others | Football News | Sky Sports,” *Sky Sports*, Jan. 07, 2016.
<https://www.skysports.com/football/news/11096/10120950/wyscout-the-scouting-tool-used-by-arsenal-manchester-united-and-others> (accessed Aug. 23, 2023).

[34]

L. Brandes and E. Franck, “Social preferences or personal career concerns? Field evidence on positive and negative reciprocity in the workplace,” *Journal of Economic Psychology*, no. 5, pp. 925–939, Oct. 2012, doi: 10.1016/j.joep.2012.05.001.

[35]

“Football League Rankings,” *Football League Rankings*.
<https://www.globalfootballrankings.com/> (accessed Aug. 23, 2023).

[36]

N. Shrestha, “Detecting Multicollinearity in Regression Analysis,” *American Journal of Applied Mathematics and Statistics*, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[37]

“Google Trends,” *Google Trends*. <https://trends.google.com/trends/> (accessed Aug. 24, 2023).

[38]

“TensorFlow,” *TensorFlow*. <https://www.tensorflow.org/> (accessed Aug. 24, 2023).

[39]

O.-C. Novac *et al.*, “Analysis of the Application Efficiency of TensorFlow and PyTorch in Convolutional Neural Network,” *Sensors*, no. 22, p. 8872, Nov. 2022, doi: 10.3390/s22228872.

[40]

Y.-Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” Apr. 2015, doi: <https://doi.org/10.11919%2Fj.issn.1002-0829.215044>.

[41]

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer Science & Business Media, 2013.

[42]

M. Tranmer, J. Murphy, M. Elliot, and M. Pampaka, “Multiple Linear Regression,” Jan. 2020, Accessed: Aug. 25, 2023. [Online]. Available: <https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf>.

[43]

P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, 2005.

[44]

V. Pestov, “Is the k-NN classifier in high dimensions affected by the curse of dimensionality?,” *Computers & Mathematics with Applications*, no. 10, pp. 1427–1437, May 2013, doi: 10.1016/j.camwa.2012.09.011.

[45]

T. Inan and L. Cavas, “Estimation of Market Values of Football Players through Artificial Neural Network: A Model Study from the Turkish Super League,” *Applied Artificial Intelligence*, no. 13, pp. 1022–1042, Sep. 2021, doi: 10.1080/08839514.2021.1966884.

[46]

E. Rey, M. Lorenzo-Martínez, R. López-Del Campo, R. Resta, and C. Lago-Peñas, “No sport for old players. A longitudinal study of aging effects on match performance in elite soccer,” *Journal of Science and Medicine in Sport*, no. 6, pp. 535–539, Jun. 2022, doi: 10.1016/j.jsams.2022.03.004.

[47]

“TransferMarkt,” *EUROPEAN LEAGUES & CUPS*. <https://www.transfermarkt.co.uk/wettbewerbe/europa> (accessed Sep. 01, 2023).

[48]

P. Garcia-del-Barrio and F. Pujol, “Hidden monopsony rents in winner-take-all markets—sport and economic contribution of Spanish soccer players,” *Managerial and Decision Economics*, no. 1, pp. 57–70, 2006, doi: 10.1002/mde.1313.

APPENDIX A – WyScout Data

	Player	Team	Position	Age	Market value	Contract expires	Minutes played	Goals	xG	Assists	...	Passes to final third per 90	Passes to penalty area per 90	Through passes per 90	Deep completions per 90	Deep completed crosses per 90	Progressive passes per 90	Co
0	Hugo Bueno	Wolverhampton Wanderers	LB	20	200000	2026-06-30	1345	0	0.14	1	...	2.14	2.54	0.33	0.33	0.94	4.68	
1	Jonny Otto	Wolverhampton Wanderers	RB, LB, RCB	29	17000000	2025-06-30	1376	1	0.27	0	...	6.21	1.70	0.39	0.33	0.33	6.87	
2	C. Dawson	Wolverhampton Wanderers	RCB	33	2500000	2025-06-30	2136	1	2.09	0	...	5.60	0.34	0.38	0.13	0.00	7.92	
3	Toti	Wolverhampton Wanderers	LB, LCB	24	2500000	2027-06-30	894	1	0.26	0	...	3.32	1.21	0.50	0.50	0.40	5.54	
4	Matheus Nunes	Wolverhampton Wanderers	AMF, LW, LDMF	24	45000000	2027-06-30	2489	1	2.85	1	...	3.36	1.95	0.54	0.40	0.40	3.76	
5	R. Ait Nouri	Wolverhampton Wanderers	LB	21	22000000	2026-06-30	1142	1	1.24	0	...	3.70	2.44	0.55	0.71	0.55	6.46	
6	João Moutinho	Wolverhampton Wanderers	AMF, LCMF, LDMF	36	2000000	2023-06-30	1944	0	1.27	2	...	6.71	2.36	1.16	0.88	0.46	6.85	
7	Diego Costa	Wolverhampton Wanderers	CF	34	3500000	2023-06-30	1232	1	3.67	0	...	0.73	0.88	0.29	0.66	0.00	0.88	
8	M. Lemina	Wolverhampton Wanderers	LCMF, LDMF, DMF	29	10000000	2025-06-30	1218	0	0.31	0	...	3.99	0.67	0.52	0.22	0.15	4.88	
9	Nélson Semedo	Wolverhampton Wanderers	RB	29	15000000	2023-06-30	2593	0	0.19	1	...	4.06	1.70	0.38	0.45	0.49	7.08	

Table A. 1 - Sample dataset from WyScout

Metric Collected	Description	Metric Type
Player	Name of the football player.	
Team	Football team the player belongs to.	
Market value	Current market value of the player.	
Contract expires	Date when the player's contract ends.	
Age	Age of the player.	Non – Performance Based
Birth Country	Country of birth of the player.	
Height	Height of the player in cm.	
Weight	Weight of the player in kg.	
Position	Playing position of the player.	Performance-Based
Minutes played	Total minutes the player has played.	
Goals	Total goals scored by the player.	
xG	Expected goals for the player.	
Assists	Total assists made by the player.	
xA	Expected assists for the player.	
Duels per 90	Average duels engaged in per 90 minutes played.	
Defensive duels per 90	Average defensive duels engaged in per 90 minutes played.	
Aerial duels per 90	Average aerial duels engaged in per 90 minutes played.	
Sliding tackles per 90	Average sliding tackles made per 90 minutes played.	

PAdj Sliding tackles	Possession-adjusted sliding tackles.
Shots blocked per 90	Average shots blocked per 90 minutes played.
Interceptions per 90	Average interceptions made per 90 minutes played.
PAdj Interceptions	Possession-adjusted interceptions.
Goals per 90	Average goals scored per 90 minutes played.
xG per 90	Expected goals per 90 minutes played.
Shots	Total shots taken by the player.
Shots per 90	Average shots taken per 90 minutes played.
Assists per 90	Average assists made per 90 minutes played.
Crosses per 90	Average crosses made per 90 minutes played.
Crosses from left flank per 90	Average crosses from left flank per 90 minutes played.
Crosses from right flank per 90	Average crosses from right flank per 90 minutes played.
Crosses to goalie box per 90	Average crosses into the goalie box per 90 minutes played.
Dribbles per 90	Average dribbles made per 90 minutes played.
Progressive runs per 90	Average progressive runs made per 90 minutes played.
Accelerations per 90	Average accelerations per 90 minutes played.
Received passes per 90	Average passes received per 90 minutes played.
Received long passes per 90	Average long passes received per 90 minutes played.
Passes per 90	Average passes made per 90 minutes played.
Forward passes per 90	Average forward passes made per 90 minutes played.
Back passes per 90	Average back passes made per 90 minutes played.
Lateral passes per 90	Average lateral passes made per 90 minutes played.
Short / medium passes per 90	Average short/medium passes made per 90 minutes played.
Long passes per 90	Average long passes made per 90 minutes played.
Average pass length, m	Average length of passes in meters.
Average long pass length, m	Average length of long passes in meters.
xA per 90	Expected assists per 90 minutes played.
Shot assists per 90	Average shot assists per 90 minutes played.
Second assists per 90	Average second assists per 90 minutes played.

Key passes per 90	Average key passes made per 90 minutes played.	
Passes to final third per 90	Average passes to the final third per 90 minutes played.	
Passes to penalty area per 90	Average passes to the penalty area per 90 minutes played.	
Through passes per 90	Average through passes made per 90 minutes played.	
Deep completions per 90	Average deep completions per 90 minutes played.	
Deep completed crosses per 90	Average deep completed crosses per 90 minutes played.	
Progressive passes per 90	Average progressive passes made per 90 minutes played.	
Conceded goals (Goalkeeper only)	Total goals conceded by goalkeeper.	
Shots against (Goalkeeper only)	Total shots faced by the goalkeeper.	
xG against (Goalkeeper only)	Expected goals against the goalkeeper.	
Prevented goals (Goalkeeper only)	Goals prevented by the goalkeeper.	

Table A. 2 - WyScout metrics and their descriptions

WyScout Position	Simplified Position
LWF	LW
RWF	RW
RCMF	CM
LCMF	CM
LDMF	DMF
RDMF	DMF
LCB	CB
RCB	CB
RAMF	AMF
LAMF	AMF
CF	CF
GK	GK

Table A. 3 - WyScout positions and how they have been simplified for this study.

APPENDIX B – Further Results

	feature	VIF
0	Goals	3.990698
1	xA	5.339883
2	Defensive duels per 90	3.169028
3	Aerial duels per 90	2.154822
4	Sliding tackles per 90	1.331534
5	Shots blocked per 90	2.760458
6	PAdj Interceptions	4.313873
7	xG per 90	9.900249
8	Shots per 90	8.381581
9	Assists per 90	2.416420
10	Crosses from left flank per 90	4.249445
11	Crosses from right flank per 90	4.591115
12	Crosses to goalie box per 90	4.249826
13	Dribbles per 90	6.143958
14	Progressive runs per 90	7.385448

Table B. 1 - VIF values after iteratively removing features greater than 10 for performance metrics.

		coef	std err	t	P> t	[0.025	0.975]
0	const	11820000.0	2.98e+05	39.613	0.000	1.12e+07	1.24e+07
20	Lateral passes per 90	5455000.0	5.79e+05	9.420	0.000	4.32e+06	6.59e+06
1	Goals	4965000.0	5.96e+05	8.328	0.000	3.8e+06	6.13e+06
29	xG against	2928000.0	5.97e+05	4.908	0.000	1.76e+06	4.1e+06
2	xA	2540000.0	6.9e+05	3.683	0.000	1.19e+06	3.89e+06
25	Key passes per 90	2390000.0	7.41e+05	3.226	0.001	9.37e+05	3.84e+06
7	PAdj Interceptions	2174000.0	6.2e+05	3.508	0.000	9.59e+05	3.39e+06
16	Accelerations per 90	2159000.0	6.18e+05	3.494	0.000	9.47e+05	3.37e+06
8	xG per 90	2049000.0	9.39e+05	2.182	0.029	2.07e+05	3.89e+06
19	Back passes per 90	1780000.0	5.66e+05	3.144	0.002	6.7e+05	2.89e+06
14	Dribbles per 90	1517000.0	7.4e+05	2.051	0.040	6.64e+04	2.97e+06
18	Forward passes per 90	1362000.0	8.5e+05	1.602	0.109	-3.05e+05	3.03e+06
10	Assists per 90	1325000.0	4.64e+05	2.856	0.004	4.15e+05	2.24e+06
17	Received long passes per 90	1113000.0	5.36e+05	2.076	0.038	6.17e+04	2.16e+06
28	Deep completions per 90	938500.0	6.56e+05	1.431	0.152	-3.47e+05	2.22e+06
27	Through passes per 90	810900.0	5.01e+05	1.620	0.105	-1.71e+05	1.79e+06
24	Second assists per 90	807800.0	3.48e+05	2.318	0.021	1.24e+05	1.49e+06
22	Average long pass length, m	450600.0	6.57e+05	0.686	0.493	-8.37e+05	1.74e+06
5	Sliding tackles per 90	281000.0	3.44e+05	0.816	0.415	-3.94e+05	9.56e+05
30	Prevented goals	68570.0	3.01e+05	0.227	0.820	-5.23e+05	6.6e+05
13	Crosses to goalie box per 90	-359600.0	6.15e+05	-0.585	0.559	-1.57e+06	8.47e+05
15	Progressive runs per 90	-389500.0	8.11e+05	-0.480	0.631	-1.98e+06	1.2e+06
6	Shots blocked per 90	-605100.0	4.96e+05	-1.220	0.223	-1.58e+06	3.67e+05
26	Passes to final third per 90	-999400.0	7.61e+05	-1.314	0.189	-2.49e+06	4.93e+05
4	Aerial duels per 90	-1028000.0	4.38e+05	-2.346	0.019	-1.89e+06	-1.69e+05
9	Shots per 90	-1138000.0	8.64e+05	-1.317	0.188	-2.83e+06	5.56e+05
3	Defensive duels per 90	-1379000.0	5.31e+05	-2.595	0.010	-2.42e+06	-3.37e+05
21	Long passes per 90	-1523000.0	7.08e+05	-2.153	0.031	-2.91e+06	-1.35e+05
11	Crosses from left flank per 90	-2135000.0	6.15e+05	-3.471	0.001	-3.34e+06	-9.29e+05
12	Crosses from right flank per 90	-2269000.0	6.39e+05	-3.548	0.000	-3.52e+06	-1.01e+06
23	Shot assists per 90	-4093000.0	8.36e+05	-4.893	0.000	-5.73e+06	-2.45e+06

Table B. 2- Full linear regression table for performance metrics

0

0	Conceded goals
1	Passes per 90
2	Crosses per 90
3	Short / medium passes per 90
4	PAdj Sliding tackles
5	xG
6	Shots against
7	Received passes per 90
8	Shots
9	Passes to penalty area per 90
10	Assists
11	Interceptions per 90
12	xA per 90
13	Duels per 90
14	Progressive passes per 90
15	Average pass length, m
16	Goals per 90
17	Deep completed crosses per 90

Table B. 3 - Dropped features due to high VIF values.

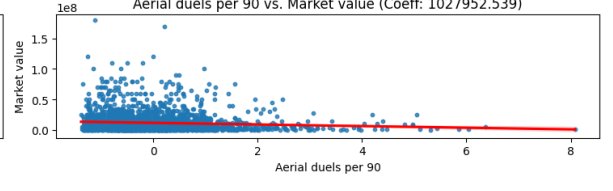
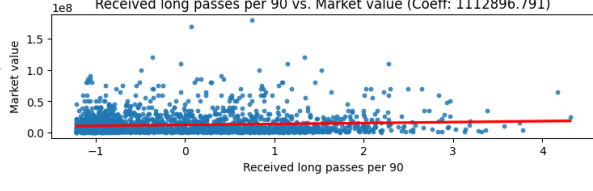
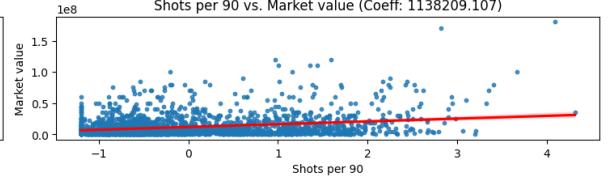
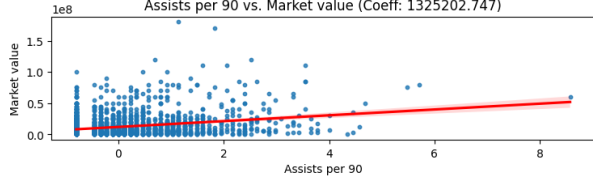
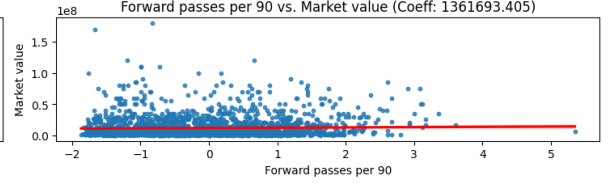
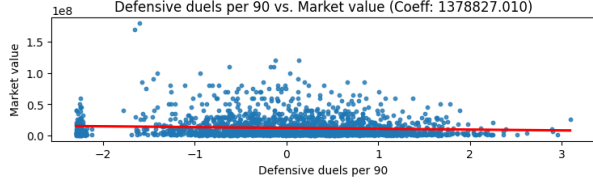
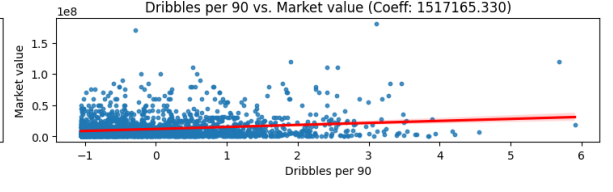
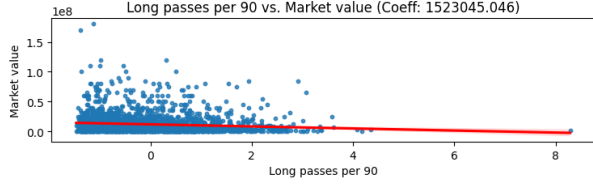
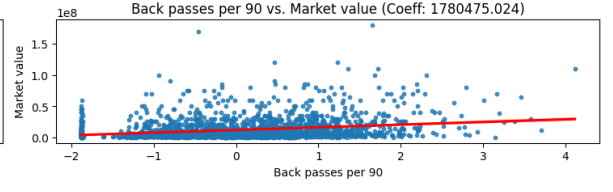
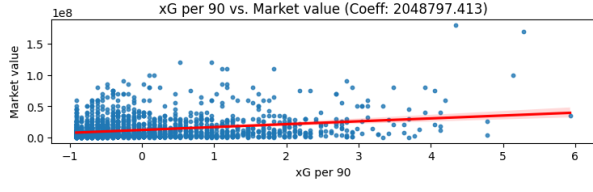
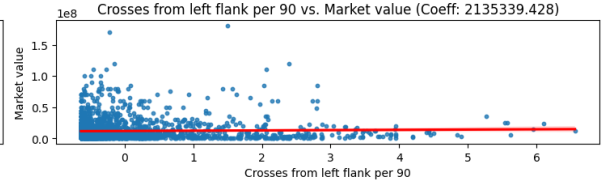
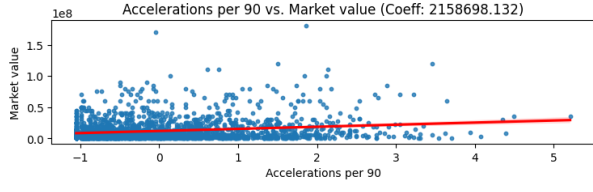
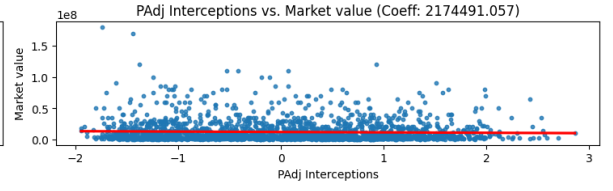
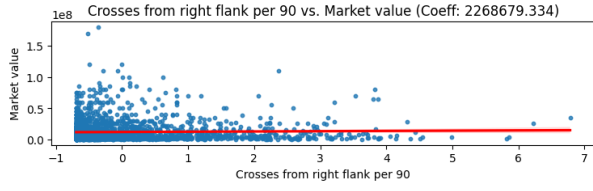
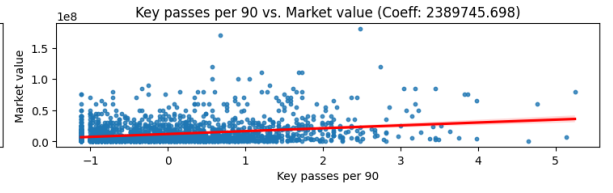
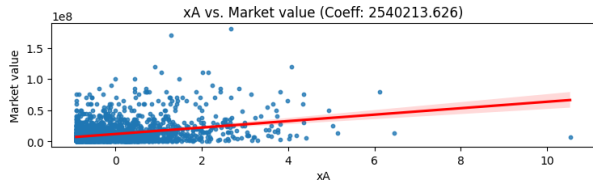
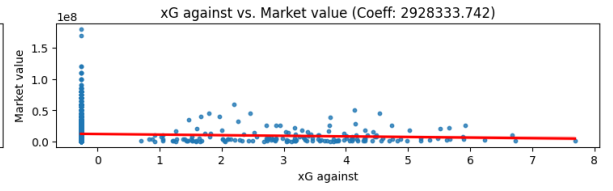
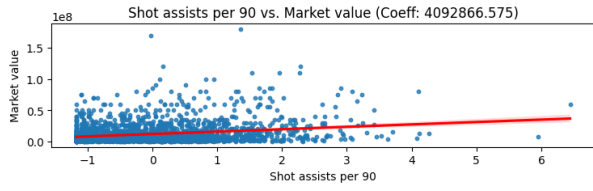
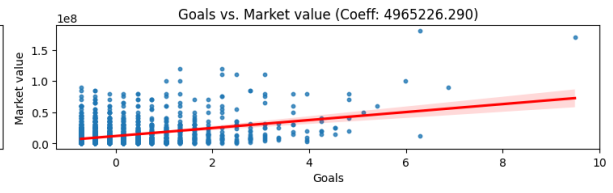
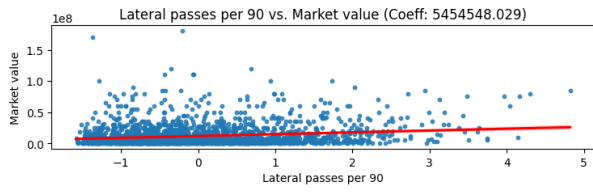


Figure B. 2- Scatter chart showing feature relationships with market value, with regression line.

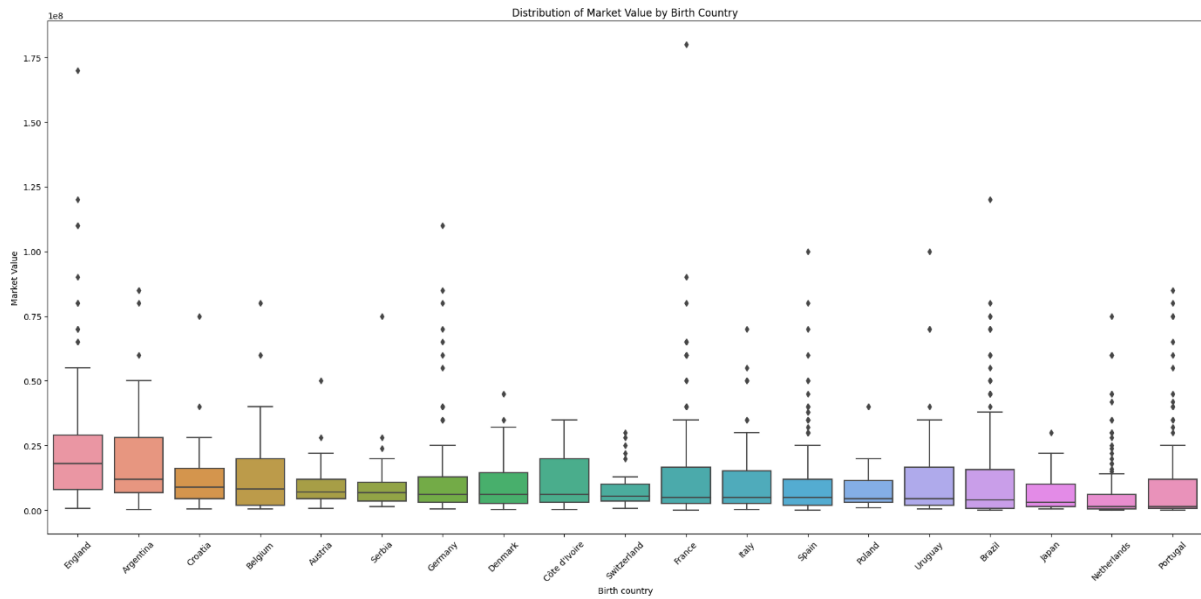


Figure B. 3 - Distribution of Market Value by Birth Country

```
# Load the Excel file
df = pd.read_excel("PlayerNames.xlsx")

# Extract player names
players = df['Player'].tolist()

# Your Google Cloud API key and Custom Search Engine ID
api_key = 'API KEY'
cse_id = 'CSE ID'

# Store the results
results = []

# Loop through player names and fetch Google search results count
for player in players:
    print(f"Fetching data for {player}...")
    try:
        query = f'{player} footballer'
        url = f"https://www.googleapis.com/customsearch/v1?q={query}&key={api_key}&cx={cse_id}"

        response = requests.get(url)
        data = response.json()

        search_count = data['queries']['request'][0]['totalResults']

        results.append({'Player': player, 'Search Results Count': search_count})

        print(f"Finished fetching data for {player}. Total results: {search_count}")
    except Exception as e:
        print(f"Error fetching data for {player}: {e}")

    time.sleep(1) # Add a 1-second delay here

# Convert results to a DataFrame
search_df = pd.DataFrame(results)

# Save the results to a new Excel file
search_df.to_excel("player_google_search_results.xlsx", index=False)
```

Figure B. 4 - Screenshot of Python code to gather search results data using Google API.

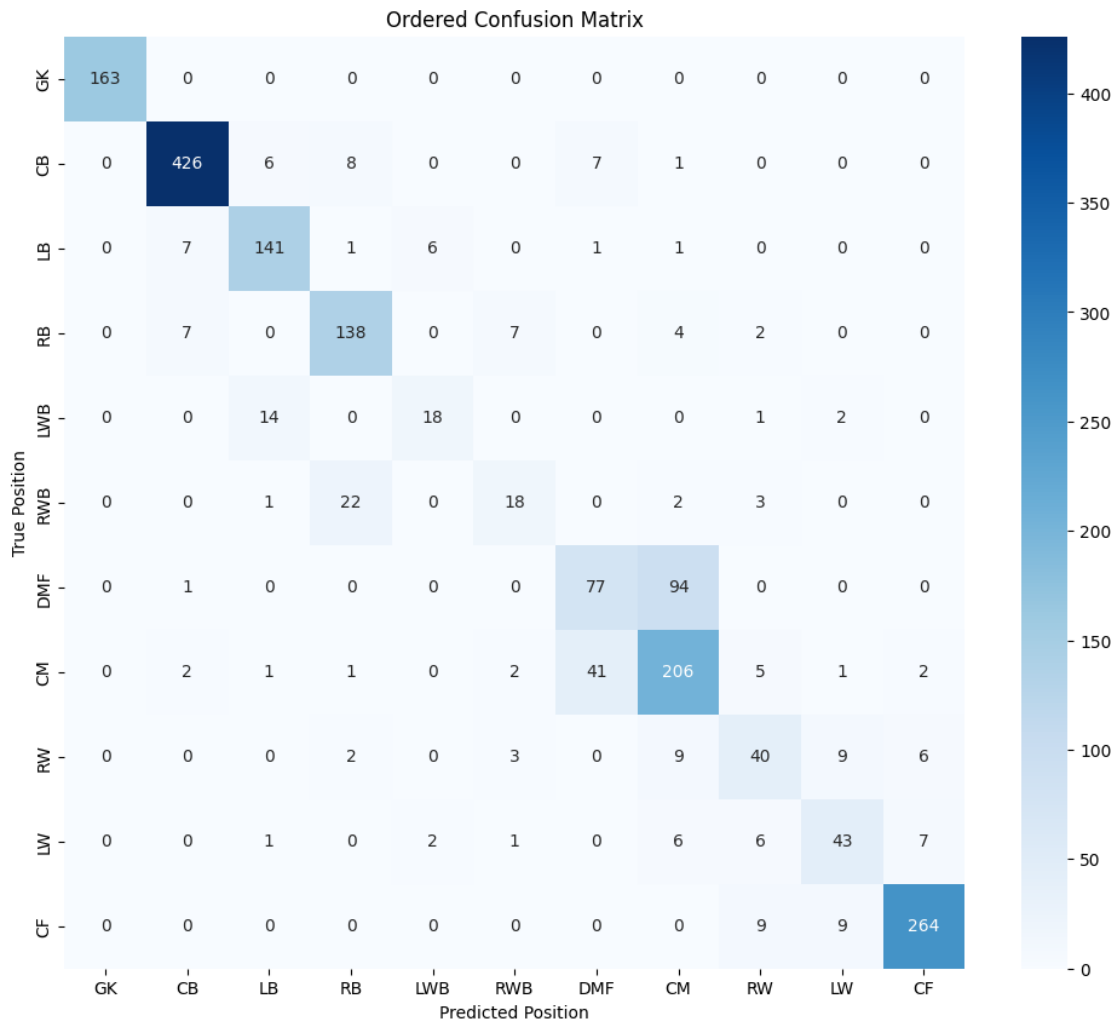


Figure B. 5 - Confusion matrix for position classification.