

***AISHWARYA SINGH***

***MSc Business Analytics***

***Candidate ID- 296644***

***WAGETASTIC***

***A Sport Analytics Study***

ACKNOWLEDGEMENT

I would like to convey my deep gratitude to Dr. Jiabin Luo for her support and detailed feedbacks over the course of this project. I am really grateful for her genuine and insightful input and direction and being extremely patient with me. Many thanks to the MSc Business Analytics team at Aston University for instilling the skills in me and availing resources that helped me complete the project. The knowledge gained during the course will be useful for the foreseeable future as well as for finishing this report. I would also like to extend my thanks to my family, friends and classmates for their unwavering support during my education.

TABLE OF CONTENT

**ABSTRACT………………………………………………………………………………………………………………………………8**

**CHAPTER I- INTRODUCTION  
1.1 HISTORY…………………………………………………………………………………………………………………………...9  
1.2 BACKGROUND……………………………………………………………………………………………………………..…..10**

1.2.1 SKILLS**………………………………………………………………………………………………………………..10**  
1.2.2 LABOUR MARKET IN PROFESSIONAL SPORTS**……………………………………………………………12**  
1.2.3 TRANSFER MARKET**………………………………………………………………………………………………12**  
1.2.4 MARKET VALUE OF FOOTBALLERS**………………………………………………………………………….13**  
1.2.5 CLUB REVENUE**……………………………………………………………………………………………………14**

**1.3 OBJECTIVE……………………………………………………………………………………………………………………….14  
1.4 STRUCTURE OF DISSERTATION……………………………………………………………………………………………15**

**CHAPTER II- LITERATURE REVIEW**

**2.1 INTRODUCTION………………………………………………………………………………………………………………..16**

**2.2 WAGE PREDICTION STUDIES………………………………………………………………………………….……………16**

2.2.1 FOOTBALL AS LABOUR MARKET**………………………………………………………………..……………16**  
2.2.2 COUNTRY BASED STUDIES ACROSS EUROPE**……………………………………………….……………19**

**2.3 CLUB REVENUE…………………………………………………………………………………………………………………20  
2.4 FURTHER ANALYSES ………………………………………………………………………………….………………………21**

2.4.1 TRANSFER MARKET**………………………………………………………………………………………………21**  
2.4.2 MAJOR FOOTBALL EFFECTS**……………………………………………………………………………………23**

**2.5 CONCLUSION……………………………………………………………….………………..…………………………………24**

**CHAPTER III- DATA**

3.1 PART A- DATA FOR WAGE PREDICTION MODEL**………………………………………………………………………26**  
3.2 PART B- DATA FOR CLUB REVENUE ANALYSIS**…………………………………………………………………………28**

**CHAPTER IV -METHODOLOGY**

**4.1 ASSUMPTIONS AND HYPOTHESES………………..……………………………………………………………………29  
4.2 WAGE PREDICTION BASED ON PLAYER PERFORMANCE AND MARKET VALUE…………………………29**

4.2.1 DATA ANALYSIS**……………………………………………………………………………………………………29**  
4.2.2 DATA PRE-PROCESSING**………………………………………………………………………………………..30**

4.2.2.1 NULL VALUES**………………………………………………………………………………………..30**  
4.2.2.2 OUTLIERS**……………………………………………………………………………………………..31**

**4.2.3 EXPLORATORY DATA ANALYSIS…………………………………………………………………………….32**

4.2.3.1 DESCRIPTIVE STATISTICS**……………………………………………………………………..….34**  
4.2.3.2 GRAPHICAL REPRESENTATIONS**……………………………………………….……………….34**

**4.2.4 DATA CLEANING AND TRANSFORMATION**

4.2.4.1 Outlier**………………………………………………………………………………………………….48**  
4.2.4.2 Feature selection**………………………………………………………………………………..…49**  
4.2.4.3 One Hot Encoding**…………………………………………………………………………….……49**  
4.2.4.4 Feature scaling**………………………………………………………………………………………49**

**4.2.5 MODELING USING MACHINE LEARNING ALGORITHMS IN PYTHON**

4.2.5.1 BASELINE MODEL**…………………………………………………………………………….……50**  
4.2.5.2 LINEAR REGRESSION**………………………………………………………………………………50**  
4.2.5.3 Decision Tree **………………………………………………………………………….……………51**  
4.2.5.4 RANDOM FOREST **…………………………………………………………………………….……54**  
4.2.5.5 SUPPORT VECTOR MACHINE**……………………………………………………………………55**  
4.2.5.6 ADABOOST **………………………………………………………………………………………..…57**  
4.2.5.7 XGBOOST**……………………………………………………………………………………..………59**  
4.2.5.8 K-NEAREST NEIGHBOUR**………………………………………………………………………….61**

**4.3 CLUB REVENUE ANALYSIS- GRAPHICAL ANALYSIS USING TABLEAU**

4.3.1 ANALYSIS OF THE TOP 10 CLUBS ACROSS FIVE MAJOR LEAGUES**…………………………………62**  
 4.3.1.1 CLUB REVENUE OF TOP 10 CLUBS**…………………………………………………………….62**  
 4.3.1.2 TRANSFER INCOME VS TRANSFER EXPENDITURE**…………………………………………63**  
4.3.1.3 RELEASE CLAUSE VS VALUE**……………………………………………………………………..64**  
 4.3.1.4 HEATMAP- WAGES PAID TO THE PLAYERS**………………………………………………….64**

4.3.2 MANCHESTER UNITED VS STOKE CITY- ANALYSIS OF TEAMS OF DIFFERENT ENGLISH LEAGUES**……………………………………………………………………………………………………………………………….65**

4.3.2.1 Manchester United (MUFC) **……………………………………………………………………65**  
4.3.2.2 Stoke City (SCFC) **…………………………………………………………………………………..66**  
4.3.2.3 KEY ANALYSIS**………………………………………………………………………………………..66**

**CHAPTER V- RESULTS AND DISCUSSION**

**5.1 PART A- WAGE PREDICTION MODEL**

5.1.1 RMSE COMPARISON**………………………………………………………………………………………….…68**  
5.1.2 FEATURE IMPORTANCE (DATASET LEVEL EXPLANATION) **…………………………………….….…70**  
5.1.3 ACTUAL VS PREDICTED WAGES**………………………………………………………………………….…..73**  
5.1.4 INSTANCE LEVEL EXPLANATIONS**……………………………………………………………………………75**  
5.1.5 SUPERSTAR EFFECT**……………………………………………………………………………………………...78**  
5.1.6 TOURNAMENT EFFECT**………………………………………………………………………………………….79**

**5.2 PART B- CLUB REVENUE ANALYSIS**

**CHAPTER VI- CONCLUSION………………………………………………………………………………………………….83**

**FUTURE WORK…………………………………………………………………………………………………………………….85**

**REFERENCES………………………………………………………………………………………………………………………..86**

**APPENDIX……………………………………………………………………………………………………………………………92**

GRAPH INDEX

|  |  |  |
| --- | --- | --- |
| **Graph ID** | **Description** | **Page number** |
| **Graph2.1** | **Player value vs salary earned in the EPL** | **19** |
| **Graph2.2** | **European clubs 20 highest transfer expenditure since 2012** | **22** |
| **Graph 4.1** | **Variables with no outliers** | **32** |
| **Graph 4.2** | **Variables with Outliers** | **32** |
| **Graph 4.3** | **To check outliers and anomalies** | **33** |
| **Graph 4.4** | **COUNTS OF PLAYER FROM COUNTRY** | **35** |
| **Graph 4.5** | **Position vs Frequency (Counts of Player per Position)** | **36** |
| **Graph 4.6** | **Best Position Values- Before regrouping** | **37** |
| **Graph 4.7** | **Best Position Values- After regrouping** | **37** |
| **Graph 4.8** | **Counts of Player Preferred Foot** | **38** |
| **Graph 4.9** | **Player Count for Week Foot** | **38** |
| **Graph 4.10** | **Skill Moves vs Player Count** | **39** |
| **Graph 4.11** | **Skill Moves vs Wage** | **39** |
| **Graph 4.12** | **Potential vs Value** | **39** |
| **Graph 4.13** | **Frequency Distribution 1** | **40** |
| **Graph 4.14** | **Frequency Distribution 2** | **41** |
| **Graph 4.15** | **Frequency Distribution 3** | **42** |
| **Graph 4.16** | **Age Distribution Of Players** | **42** |
| **Graph 4.17** | **Potential Of Players Distribution** | **43** |
| **Graph 4.18** | **Clubs vs Player's Average Value** | **43** |
| **Graph 4.19** | **Player Based On Value** | **44** |
| **Graph 4.20** | **Release Clause vs Value** | **45** |
| **Graph 4.21** | **Name vs Release Clause** | **45** |
| **Graph 4.22** | **Age vs Value** | **46** |
| **Graph 4.23** | **Wage vs Age** | **46** |
| **Graph 4.24** | **Wage vs Overall** | **47** |
| **Graph 4.25** | **Heatmap- Attributes correlation** | **48** |
| **Graph 4.26** | **Actual vs Predicted (first 50 instances)** | **51** |
| **Graph 4.27** | **Club Revenue of Top 10 clubs** | **63** |
| **Graph 4.28** | **Transfer expenses of top 10 clubs** | **63** |
| **Graph 4.29** | **Release Clause vs Value** | **64** |
| **Graph 4.30** | **HEATMAP- Highest Wage paying Clubs** | **65** |
| **Graph 4.31** | **Club revenue of Manchester United** | **66** |
| **Graph 4.32** | **MUFC vs SCFC revenue comparison** | **67** |
| **Graph 5.1** | **Actual vs Predicted (ADA Boost)** | **70** |
| **Graph 5.2** | **Feature importance** | **71** |
| **Graph 5.3** | **MUFC- Actual wage vs Predicted wage against Reactions rating** | **72** |
| **Graph 5.4** | **SCFC- Actual wage vs Predicted wage against Reactions rating** | **72** |
| **Graph 5.5** | **MUFC- Actual Wage vs Predicted wage comparison** | **73** |
| **Graph 5.6** | **SCFC- Actual Wage vs Predicted wage comparison** | **73** |
| **Graph 5.7** | **MUFC- Wage distribution based on Position** | **74** |
| **Graph 5.8** | **SCFC- Wage distribution based on Position** | **75** |
| **Graph 5.9** | **International Reputation MUFC vs SCFC** | **76** |
| **Graph 5.10** | **SCFC salary distribution based on player position over time** | **80** |
| **Graph 5.11** | **MUFC salary distribution based on player position over time** | **81** |
| **Graph 5.12** | **EPL Clubs tendency to pay inflated salaries** | **82** |

TABLE AND FIGURE INDEX

|  |  |  |
| --- | --- | --- |
| **Table ID** | **Description** | **Page number** |
| **Table 3.1** | **Wage Prediction Dataset description** | **27** |
| **Table 3.2** | **CLUB REVENUE ANALYSIS DATASET** | **28** |
| **Table 4.1** | **Variables removed** | **30** |
| **Table 4.2** | **Null Values** | **31** |
| **Table 4.3** | **Top 5 values after removing text** | **33** |
| **Table 4.4** | **Descriptive Statistics for fields with money details** | **33** |
| **Table 4.5** | **Descriptive Statistics of all the variables** | **34** |
| **Table 4.6** | **COUNTS OF PLAYER FROM COUNTRY** | **35** |
| **Table 4.7** | **COUNTS OF PLAYER PER POSITION** | **36** |
| **Table 4.8** | **Player Based On Value vs Overall and Position** | **44** |
| **Table 4.9** | **Top 15 Release Clauses** | **45** |
| **Table 4.10** | **Best\_Position regrouped** | **48** |
| **Table 4.11** | **RMSE on Hyperparameter combination- Decision Tree** | **53** |
| **Table 4.12** | **RMSE on Hyperparameter combination- Random Forest** | **55** |
| **Table 4.13** | **RMSE on Hyperparameter combination- SVM** | **56** |
| **Table 4.14** | **RMSE on Hyperparameter combination- ADA Boost** | **57** |
| **Table 4.15** | **RMSE on Hyperparameter combination- XG Boost** | **60** |
| **Table 4.16** | **RMSE on Hyperparameter combination- KNN** | **61** |
| **Table 4.17** | **Manchester Utd vs Stoke City revenue and expense data** | **66** |
| **Table 5.1** | **RMSE Comparison** | **68** |
| **Table 5.2** | **SCATTERPLOTS (PREDICTED VS ACTUAL)** | **68** |
| **Table 5.3** | **Case 5- MUFC players comparison with same OVR** | **76** |
| **Table 5.4** | **CASE 6- SCFC- Same Rating- 71** | **77** |
| **Table 5.5** | **Jones vs Owen** | **77** |
| **Table 5.6** | **Owen vs Ronaldo** | **77** |
| **Table 5.7** | **MUFC Players Sample** | **78** |

|  |  |  |
| --- | --- | --- |
| **Figure ID** | **Description** | **Page number** |
| **Figure 1.1** | **World map showing top FIFA national football teams** | **9** |
| **Figure 4.1** | **Decision Tree** | **52** |
| **Figure 4.2** | **Hyperplane in 2D and 3D** | **55** |
| **Figure 4.3** | **Ada Boost model** | **57** |
| **Figure 4.4** | **XGBoost Model** | **59** |
| **Figure 5.1** | **Club Revenue of Top 20 clubs** | **80** |

ABSTRACT

We examine the factors that affect player wages in professional football in two steps using a FIFA dataset with 16710 players from more than 80 different countries and 54 attributes describing them, across various leagues and divisions. A major part of the research focuses on modeling a predictive algorithm for calculating wages. We discover that individual incomes are influenced by OVR, value, reactions, international reputation and ball control. Additionally, we show how the effects of these factors differ across the income distribution. ADABoost proved to be the best model with RMSE of 11064 on the test set. The suggested approach can be utilised to do quantitative analyses of the relationships between football players' salaries and performance as well as assist in negotiating player salaries. The approach is not only based on the performance and abilities of the players, but it also accounts for indirect factors related to the game, such as the player's value and international reputation as they also have a significant impact on the salary, particularly in the case of star players and team captains. Major football effects such as superstar and tournament effects helped in explaining the gaps in Chapter 5. The second part briefly explains how the club revenues of the five top leagues across Europe and different division leagues within England affect the football market and players’ wages. The analysis was performed using graphical visualizations on the data collected manually from the annual revenue reports of the clubs and transfermarkt. The operations of the top leagues in terms of overspending and monopolising the market seemed consistent whereas the lower leagues face dearth of talents due to less revenues.

**Keywords:** Football; Wages; Machine Learning; Club Revenue

CHAPTER I  
INTRODUCTION

**1.1 HISTORY**

Football stretches back to between 200 and 300 BC, making it one of the oldest games in the world, according to FIFA (Wood, 2008). Back in the 9th century, Britain called it the famous game of ball played by the then youths of London from various social classes, every Tuesday with minimal equipment as opposed to the extravagant vocations. It was infamous for instigating riots and deaths and was deemed to be a dangerous sport especially during the English Civil wars. Despite being enjoyable, the game was outlawed numerous times before the 18th century, including by a decree in 1409 that prohibited play by labourers. It was not until 1741 that judges were officially appointed to resolve chaos and the rules that are recognized today started taking shape. Matches in factories and public schools were organized and a more civic version of football was developed (Hessayon, 2014). The late 19th century witnessed the dawn of Association football and local clubs’ rivalries. The clubs were constituted majorly of factory workers and other workforces and they started building on the local community support. This brought about professionalization in football and the players started emerging as heroes of their respective clubs. Various countries came to the forefront with their major teams and thus began the series of league matches. FIFA, the international governing body of association football, futsal and beach soccer, was founded on 21 May 1904 (FIFA, 2022). It is one of the world's oldest and largest NGOs that spread the awareness and sportsmanship of football to the other parts of the world forming 5 major confederations in all the continents.



Figure 1.1 World map showing top FIFA national football teams (Source- Alamy Stock Photo)

Football has evolved drastically over time from being a gruesomely violent peasant pastime to becoming a multi-billion dollar industry. One of the most popular sports of all time; it has more than 5 billion fans and viewers worldwide. The most commercially successful sport of today’s generation, it has garnered sponsors' benefits and branding through television rights, superstar players, merchandising, etc. Players have emerged as celebrities with agents of their own. Clubs possess their channels, group of analysts, world-class managers, and medical units. Club revenues depend on match tickets, viewership, and advertisements (Hessayon, 2014). The competition among the teams and players has given rise to some of the world-class players of all time such as Diego Maradona, Pele, Lionel Messi, etc. With the culmination of club culture and rivalry, every club necessitates having the best players in their clubs and one way to acquire such players is through transfer windows. With football becoming expensive, the transfer prices have seen an exponential increase in its value. In 1984, the transfer of Diego Maradona, the best player of his time, to Napoli cost around £6.9 million. While, the 2022’s transfer of France's Kylian Mbappé had the highest transfer value out of any player at the tournament, being worth approximately £140 million (Transfer Markt., 2022).

The buying prices, transfer prices, and the title of the player as an elite player affect the wages of the football players. Even better-performing players look for richer clubs to earn recognition, wages, and fame. Club revenue also becomes a deciding factor if a player can be afforded by the club both in terms of buying and loaning. Increasing club revenues and having and sustaining an elite player thus logically form a mutual event for any club. Therefore, small clubs are almost all the time stuck in a catch-22 situation where they need elite players to increase their club revenues but they also need high club revenues to pay such players. This kind of behaviour needs to be studied and validated across different leagues to validate if the skills of a player directly transform to higher wages or if are there other factors including and apart from club revenues that play a major role.

**1.2 BACKGROUND**

**1.2.1 SKILLS**

Football skills are quite diverse especially considering all the playable body parts of the footballers but specifically the skills can be summarized into four namely, Technique, Game, Physical fitness and Proper mindset.

**TECHNIQUE -** Elaborating on the Techniques, the skill set could be further divided into Ball control, Dribbling skills, Passing accuracy, Body control.

**1.** **Ball control-** A player's ability to control the ball while using all of their playable body parts is referred to as ball control. Starting from acquiring the ball to maintaining possession, a player with good ball control can receive passes both on the ground and out of the air (Ertheo, 2022). It also refers to player's capacity to successfully keep the ball in his or her hands while successfully defending it from opponents. Football success also heavily depends on the player's ability to react fast with the ball, which is an integral component of ball control (Desai, 2022).

**2. Dribbling skills-** Dribbling talents are the capacity to manoeuvre about the field while maintaining perfect control of the ball. A player with good dribbling ability can use both the feet while playing, in various directions and at various speeds. They can pass through opponents with ease and keep the ball in their control. For players in all positions, having excellent dribbling techniques is crucial to success in football (Desai, 2022). .

**3. Passing accuracy-** The ability to kick the ball with both feet to the intended destination efficiently is referred to as passing accuracy. Passing entails crossing the ball accurately in front of the goal or, for attackers, accurately and powerfully finnishing. It could also mean putting the ball straight to a teammate's feet with strength and precision (Desai, 2022).Without the ability to pass accurately with both feet, players cannot succeed in football.

The two most important components for any football player's success are passing and receiving. Accuracy is essential when passing and receiving. It permits consistent field movement, which may create up scoring opportunities, and smooth open play. It may be observed when watching top leagues throughout the world that the ball doesn't actually stop after it is received. It is continually moving in the direction that is required to develop play, according to Chhangte (Desai, 2022). Therefore, receiving the ball correctly in the first place, lets the player assess the alternatives leading to a variety of passes. A flawless transition can result in beneficial offensive circumstances.

**4. Body control-** The capacity of a player to move their bodies flexibly to improve balance and coordination is referred to as body control. Body control mostly relates to proper form because it falls under the realm of technique rather than physical fitness. A low centre of gravity, long strides, and proper running technique are all signs of good body control. Apart from these some of the specific skills are heading, shooting, touch, etc.

**GAME INTELLIGENCE -** The capacity of a player to act swiftly and wisely when making judgments on the field is a key indicator of game intelligence. A shrewd footballer wants to play as efficiently as possible, using their energy efficiently to achieve their goals. In essence, playing intelligently is trying to win rather than trying to lose. It comprises of skills such as Spatial awareness, Tactical knowledge, Risk assessment

While spatial awareness and tactical knowledge are inherent to being a footballer, risk assessment is a skill that comes with practice and experience. A player's capacity to perceive space clearly on the full field and use it to their advantage is referred to as **spatial awareness** (Ertheo, 2022). Players with tactics are constantly aware of where their teammates are and where they should be based on how the opposing team is positioned. This is useful when moving quickly and instinctively, almost without looking (Ertheo, 2022).

The term "**tactical knowledge**" describes a player's understanding of the rules and structure of the game, like as how a team is formed, which has a significant impact on the strategies they will employ to win.

In football, **Risk** is the possibility that certain bold choices could result in the failure to regain possession in the case of diving or tackling on defense. Every team member must work harder when they lose control of the ball until they reclaim it. Such aggressive choices could result in excessive energy expenditure and a failure to perform successfully for the entire 90 minutes of the game. On the other hand, if a team makes no aggressive decisions at all in an effort to preserve energy and keep possession, the team may not be able to score goals and may ultimately lose (Ertheo, 2022).

**PHYSICAL FITNESS** – Physical fitness comprises of abilities such as Endurance, Balance and coordination, Speed, Strength and power. Football matches typically last 90 to 95 minutes with very few substitutions, thus physical condition is another crucial component to success. According to studies, midfielders run an average of more than 11 km per game. The most "high-intensity" runs are made by wingers, who average roughly 150 sprints at least 75% of their maximum speed per game. Center-backs run the least in terms of distance travelled per game—aside from goalkeepers—averaging 9.5 km (Ertheo, 2022).

Football players need balance and physical strength in addition to endurance and speed to protect the ball at their feet, shoot, pass the ball far, win balls out of the air, etc.

**PROPER MINDSET -** The final element to success in football is having the right mindset. Footballers must dedicate their entire being to their sport in order to succeed. Having said that, such fervour can cause destruction in the event of failure or fatigue. Football players need to establish a balance, develop passion, keep their calm, and exhibit resilience if they want to succeed in the sport. Coachability, Self-motivation, Composure and mental strength, are the major areas of development that footballers need to consider before even considering the sport (Ertheo, 2022).

**1.2.2 LABOUR MARKET IN PROFESSIONAL SPORTS**

The concept of labour market has become quite prominent in professional sports and has paved way for various economic analyses. One of the first such thorough analyses, by Simon Rottenberg in 1956, focused on the labour markets in professional sports and ultimately established the general direction of much future studies. Although sport is a fantastic setting for applied economics, assessments are often made more difficult by novel and intriguing developments in technology and institutional structures. Rosen and Sanderson (2001) have analysed various issues in sports economics using conventional demand, supply, and market equilibrium methods.

**1.2.3 TRANSFER MARKET**

When a player on contract with one club goes to another, it is known as a transfer. The term "transfer" is used because the player's registration information moves from one association football club to another. The selling clubs ask for "transfer fee", which is paid by the buying club in order to make up for the selling club losing the player and their services before the end of the contract. (Football Stadiums, 2022)..

Analysing the transfer market in popular sports such as Football and Baseball is a convoluted affair as football clubs have spent astronomical transfer prices for their players. The transfer of Barcelona’s forward, Neymar to Paris St-Germain, in 2017 for a world record 222 million Euros bears testimony to the burgeoning transfer prices (Football Stadiums, 2022).

When a player is "placed on the transfer list," his club has made him available for transfer. After that, other clubs are free to approach the player's current club and make a bid to sign him. Clubs can contact other clubs to make an offer for a player, but they are aware that a player on the "transfer list" can be bought for a lower price because the club is prepared to sell the player. A player may also submit a "transfer-request" if they want to leave their team before the term of their contract expires (Iterpro, 2022).

**1.2.4 MARKET VALUE OF FOOTBALLERS**

The estimated sum for which a club or team can sell the player's contract to another is known as the market value of a football player. The transfer value is directly proportional to the contract length. From a logical perspective, following are the different factors that influence the market value of footballers.

**Age -** Younger players typically cost more than older ones. A player's value or worth decreases as they tend to retire since they have fewer years left to maintain their level of play, whereas a young player is seen as a long-term investment. Growing young players, grooming them and then selling them to a bigger club for a higher price is how many clubs generate a large portion of their revenue. Determining the market value of football players is not an exact science (James, 2022) because certain players, like Jamie Vardy, have excelled later in their careers, but as a general rule, it is very consistent.

**Contract Duration -** The transfer cost is directly proportional to the number of years a football player still has left on his or her deal with the club. Due to the fact that buying a player essentially means buying the remaining time on their contract, the more money a player is worth. In other terms, a football player's market value is a ballpark figure for how much a team may get for the rights to the player's contract from another team (James, 2022). The transfer value increases with contract length.

**Position -** Position on the pitch is a third aspect that affects a football player's market value. Attacking players and those who contribute frequently to goals or assists are frequently more expensive. Strikers are often more valuable than defenders or goalkeepers, which might not seem just. The market worth of a football player is influenced by the player's primary position on the pitch. The market value of players at various positions is determined by a variety of factors, but generally speaking, the higher a player is placed on the pitch, the more expensive he will be (James, 2022).

**Skills -** Talent is, of course, the basic basis of the game, and the better a player is, the more valuable he will be to a team. Due to Neymar's great level of skill, the Brazilian international's 222 million Euro transfer from Barcelona to Paris Saint-Germain is so costly. Players with more skill will incur higher transfer fees. According to research, players with higher average ratings typically cost more to transfer (James, 2022).

**League -** Another key factor in establishing a football player's market worth is the difficulty of the league. A player will be rated more highly if he or she is competing in one of the "big five" European leagues, such as the Premier League, Bundesliga, or La Liga. A player's market worth rises if they have previously established themselves in a competitive league because there is less risk involved in the purchase (James, 2022). That holds true for both the team selling the player and the club hoping to sign him. Selling club with the knowledge of available resources to the buying team may thus demand a higher price for their players, and consequently bigger clubs will have to spend more on players. The market worth of the player will be higher, for instance, if a club has earned a spot in the Champions League. This issue affects both the team that wants to buy the player and the team that wants to sell him.

**Brand -** The market value of a player is heavily influenced by their talent as well as their fame and "superstar status." In other words, regardless of how talented they are, football players' market value also depends on their ability to draw large crowds. The quantity of jerseys sold and money made from portrait rights are both influenced by a player's image outside of the football field. You are worth more if you have a large number of Instagram followers (James, 2022).

A club should cherish its athletes' popularity since it has economic value. Players like Messi, Ronaldo, or even Ibrahimovic are nearing retirement, but their brand worth is still very high since they have established a global reputation over the course of their careers. Everyone is familiar with their face, which offers them an advantage when negotiating sponsorship agreements with well-known commercial brands.

**1.2.5 CLUB REVENUE**

The football industry's financial landscape is highly competitive, intricate, and dynamic. Football teams have become overly dependent on TV revenue, but this is insufficient nowadays (Iterpro, 2022). The club revenue is divided into the following three categories: match day, broadcasting, and commercial:

**• Match day revenue**, which is the money earned by clubs from hosting games at their stadium and is primarily derived from ticket sales.  
• **Broadcasting revenue** is the money clubs receive from media broadcasts as a result of their involvement in domestic leagues, domestic cups, and, in some cases, international events.  
• **Commercial revenue**, which refers to the money that sports teams make from sponsorships, merchandise, and other business ventures.

Additionally, clubs also have contribution to their revenue through the players’ loans and transfers. The tournament and championship money that the teams receive for participating and winning also get added to the revenue.

**1.3 OBJECTIVE**

There are many estimations and models provided by economists, analysts and mathematicians on football analytics. But one of the major issues is the burgeoning wages of football players especially in the European with almost no caps. Not only this, but also income inequality within the clubs has become a concern. From a behavioural point of view, increasing salaries of players improves player performance (Torgler and Schmidt, 2007) while the efforts of players have proved to be insignificant when it comes to determining the salaries (Wicker et al., 2013). Some researchers have also suggested that increasing player spending does not result in improved performance and financial success (Sloane, 1971). With such contrasting opinions and issues on the spectrum, it becomes important to identify a model to determine salaries of the players and identify parameters that majorly affects it.

It is not just the players who suffer from pay inequality. Even small clubs face dearth of talents and sponsorships due to the absence of wage cap regulations and therefore do not have a chance to compete fairly. With the emergence of data science and Sports analytics, data required to carry out analysis on Players’ wages has become easily accessible. Salaries of players being published online foster pay transparency and studies on pay equity. The data collection has become immensely comprehensive in the past couple of decades with FIFA and UEFA measuring players’ performances and values almost precisely through various measures and publishing online.

There has been only a couple of wage prediction studies published in English. Therefore, using machine learning algorithms, the aim of this study to predict the wages of the players across all the leagues associated with FIFA based on various attributes. With a best model identified, the attributes that contribute the most to the target variable “Wage” will be used to explain the results through instance based explanations. Additionally, club revenues across some leagues will be studied with respect to the wages to establish the uniform and contrasting trends among clubs and across leagues. Although not elaborately, but the presence or absence of tournament and superstar effects will also be determined. The study aims at explaining what impacts wages the most- types of skills (fitness, tact, etc.), brand and market value of players (age, league, position, etc.), transfer value, club revenue, etc. This is due to the fact that the club-specific studies conducted till now have had different results such as attributes that affect the wage most, if the market is competitive, if players should be paid more, etc. and hence cannot be applied to all the leagues. Therefore, all throughout it has been my motive to present a unified study with 16710 players across the world to present a standard model.

**1.4 STRUCTURE OF DISSERTATION**

Starting from **literature review**, Chapter 2 provides the details of some of the research works that I have gone through specifically in the fields of Wages prediction, transfer market, football as labour market, club revenue and different types of effects, using machine learning, cost estimations, profit functions, etc., in order to build my understanding and awareness of the studies those already exist and those further needed.   
Followed by Chapter 3, **methodology** describes the current research in two parts. The first part talks about the “Wage prediction using ML algorithms (Python)” and the required data collection, data pre-processing and exploratory data analysis followed by various descriptive and predictive models and analyses in detail.   
The second part briefly explains the impact of club revenue on wages across five major European leagues from different countries and between two different teams of two different leagues within the same country. The analysis was conducted graphically using Tableau.  
Chapter 4 summarizes the result from both the analyses in terms of error comparison and instance based explanation. It also highlights the features those define the wages primarily. Further discussions and comparison with real time have also been conducted.  
Chapter 5 is the **conclusion** part with the analyses and the whole dissertation mapped with the existing scenario followed by **Future work** which emphasizes on the areas to be considered for further analyses and **References**.

CHAPTER II  
LITERATURE REVIEW

**2.1 INTRODUCTION**

The salary prediction has always been studied from the perspective of players’ performances and individual attributes only. A study that could encompass the effect of the club's revenue from different income sources has been missing. Such a study can aid the negotiation process where both the club management and agents can rationally understand the salaries to be discussed. Also, studies conducted until now have been solely based on the performance and skills of the player. They do not take factors such as players' popularity, brand value, etc. into account which definitely have a high impact on the players' salaries as well as the club's revenue. Also, detailed analyses have always been conducted at national levels which fail to give an encompassed and comprehensive understanding of the leagues of a given country when compared with another country’s league (Section 4.2).

Therefore, the dissertation was based on analyzing the wages of the football players from two major perspectives. The first set of literature explains the salaries of the players using various machine learning techniques using the performance set of the players and including but not limited to their popularities and market values. It also gives an insight into the various studies conducted across the major football leagues under UEFA. The second set of literature reviews the effect of the club revenue from various sources on the salary negotiation. The transfer market and the superstar effect have been finally used to shed light on the results gained from the two major analyses.

The European Court of Justice's Bosman ruling of 1995, freed players from club restrictions once their contracts expired, fostering a competitive transfer market and driving up transfer fees and salaries. Following that, transfer fees and wages gradually reached record breaking highs on a yearly or biannual basis in order to fuel the talent acquisition war that propelled football clubs' winning maximisation goals (Ezzedine, 2020, pp. 2-10). Therefore, the study to be conducted will not just estimate a player's worth but will also provide an estimate of what the clubs can afford to pay.

**2.2 WAGE PREDICTION STUDIES**

Wage prediction studies have been conducted individually in different countries but with only one country at a time. There are separate studies on Bundesliga, La Liga, etc. but unified studies are missing. This section contains the major concepts important to know before analyzing wages and their corresponding researches.

**2.2.1 FOOTBALL AS LABOUR MARKET**

An exceptional potential for labour market research exists in professional sport. But professional football players' labour markets have unique characteristics that make them difficult to study (Brocard and Lepetit, 2018). When attempting to analyse the link between supply and demand in this specific market, the application of traditional economic theories is in fact insufficient. In this market, it is particularly rare to see the standard assumptions intended to explain traditional labour markets (Brocard and Lepetit, 2018). With frequent institutional reports offering authoritative data and a well-documented market, research is made easier nowadays. An alternative theory—the segmentation of the labour market—seems to be supported by an empirical investigation of the labour market for professional football players. In fact, the latter theory contends that there are several markets or market sectors with significantly different players' behaviours and supply and demand adjustment processes rather than a single, homogenous labour market. Additionally, a historical examination of this labour market reveals that it is highly sensitive to the degree of specific regulation. Numerous deregulation choices have been taken in the previous 20 years, either by the sports industry itself or by national or international governmental authorities, which have had a significant impact on how the market functions. The significant increase in player movement internationally and the talent concentration in the top leagues and wealthiest clubs are two of the key effects.

Brocard and Lepetit (2018) stress that, value is determined by many factors than just performance numbers. Production in the entertainment industries, of which sports make up an increasing portion, is labor-intensive, and it is difficult to separate the finished product from the individuals who provide the service. Economic "output" frequently consists of an accurate assessment of the inputs themselves. Some famous sportsmen gain personal fan bases that extend far beyond their contributions to the calibre of certain tournaments. Even while "star quality" is frequently elusive and difficult to discern from straightforward numbers, fans seem to be able to spot it when they see it. It may have a significant impact on sports revenue (Hausman and Leonard, 1997).

Numerous studies have been conducted both formally and informally on the football leagues all around the world as the sports industry has taken a sharp turn towards being the biggest business market. Majority of these studies have been conducted using econometric approaches. Studies on transfer market and player position are some of the very popular studies. But literatures around the studies to determine ways of predicting the wages of the football players are limited as researchers have focused mostly on the transfer market and income inequality. Almost all the studies have had a common assumption of the **football labour market** considered in their research to elaborate on the findings later on.

Football player compensations were based on more qualitative assessments in the pre-information era when obtaining football player statistics and data was challenging (Frick, 2006). This made evaluating the abilities and performances of football players incredibly challenging. Due to unavailability of data at all the club levels and changes made to the football rules in the 1990s, Szymanski and Smith (1997) presented an empirical model from 1974 to 1989 using econometric approach. Football talent is purchased on a player market that is competitive, and the quality-adjusted wage is established as Nash equilibrium between the clubs*. A club's standing in the League is determined by the quantity of skill it purchases* (Szymanski and Smith, 1997). This was consistent with the estimated industry production function. The amount of money a team makes through gate revenues, television rights, sponsorships, etc. depended on where they were in the League, and this correlated to the industry's demand function, which was also evaluated by Szymanski and Smith (1997). They derived the empirical trade-off between profit and league position that the club faced by fusing the estimated demand and production functions with the budget restriction and finally revealed that *increasing player spending does not result in improved performance and financial success*. Profits and status both play a role in the owner of the club's objective function (Sloane, 1971). The ideal wage level, as well as the club's profits and League standing, were determined by maximising the objective function under the profit-position constraint. Profit and position were negatively correlated inside clubs, while they were positively correlated across clubs, so each club must choose between the two. This is due to the fact that clubs' endowments, inherent capacities, and capacity for generating money vary, and these endowment variations were what produced the observed distribution of profit and position.

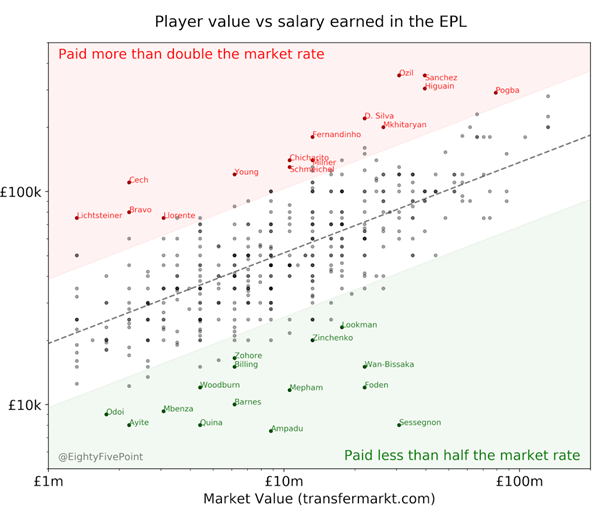
The compensations could be easily estimated and compared in more recent years, notably since the 1990s when statistics and data about football players were easily accessible. Indeed, Frick (2006) underlined that player statistics and data were made public and that comparisons at many levels, including several performance metrics and wages, were possible because of the data's comprehensiveness. Several more facts regarding football players have been made public and have been proven to be helpful in determining player compensations*:* ***a player's innate qualities, aptitude, and human capital* (Szymanski and Smith, 1997); *additional revenue generated for the team through ticket sales, merchandise sales, and broadcasting deals* (Garcia-del Barrio and Pujol, 2007); *a player's performance in the previous season, games played abroad, and the number of goals scored by the player* (Frick, 2011); *passing accuracy, free-kick speed, and tackles;* and other factors (Yaldo & Shamir, 2017).**

Since the data was collected on an annual basis rather than match per match, Szymanski and Smith (1997) did not make an effort to account for the influence of outcome uncertainty, which has been demonstrated to affect demand. As a result, it made more sense to analyse player salaries using data-specific models. One such study on computing the football wages by Yaldo and Shamir (2017), focused entirely on predicting the wages based on the footballer’s performance and skill dataset. The study accounted for variables such as football skills and general physiological variables, and aimed to explain how transfer prices affected the players’ wages. The system outlined in the paper was driven by quantitative analysis in order to further justify the behavioural impact of the negative impact of the salary inequality on the team and the difference between skills of overpaid and underpaid players. Also, the supervised machine learning methods implemented in the study covered eight different pattern recognition algorithms such as Nearest Neighbor with a weighted condition (Aha et al., 1991), Additive Regression (Friedman, 2002), Decision Table (Kohavi, 1995), Random Trees (Aldous, 1993), etc.

But the mean absolute error values of the models were varied and the predicted output was far from the actual output. For example, the actual weekly salary of Harry Kane was €15,000 whereas the predicted salary was €119,798 in 2016. The study successfully identified the top 100 highest and 100 lowest paid players but the salaries predicted using pattern recognition machine learning algorithms were not very accurate. A simple explanation is the fact that the author didn’t account for the factor known as the “superstar effect” which did not form the basis of the research for the paper. Additionally, the author recommended that ***the wage of a football player is determined by several factors that are not directly related to performance or talents***. For instance, a player who is well-liked by the audience may receive more pay to account for merchandising and ticket sales.

**2.2.2 COUNTRY BASED STUDIES ACROSS EUROPE**

The English Premiere League is a long-standing industrial cartel that sells a very well-liked good with only mediocre alternatives. Despite this, the majority of its member clubs are in the red, and the business has had numerous financial problems (Szymanski & Smith, 1997). The amount of money available to football clubs has increased due to growing competition among television networks. The disruption of the pre-existing terrestrial "duopsony" by satellite broadcasting in particular caused a tenfold increase in broadcast income. The most prominent result of this alteration was that the top clubs formed the Premier League and split off from the rest of the Leagues, keeping practically all of the television money for themselves (Szymanski & Smith, 1997).



Graph2.1 Player value vs salary earned in the EPL

Source: eightyfivepoints

Information on 533 outfield players from the Italian "Serie A" and "Serie B" before the start of the 1995-96 season is used by Lucifora and Simmons (2003). They discover that the *number of games played and goals scored*, which are the primary indicators of individual performance, *have a statistically significant and economically substantial impact on compensation*. Additionally, earnings are strongly correlated with a person's career goal-scoring rate and assist rate, indicating a sizable "superstar effect." 651 player-year observations from the German "Bundesliga" for the seasons of 1998–1999 and 1999–2000 are used by Lehmann and Schulze (2008). Additionally, their performance metrics have a predictable and statistically significant impact on wages. The concept of "superstardom" was difficult to reconcile with the surprising discovery that media presence has a positive, but reducing influence, implying decreasing returns to popularity. Au Contraire, Frick (2008) showed in his study the existence the of superstar effect.

The study on the salary of football players highlighted the shift in public view from 1954 to 1962 when the concern for the huge differences in the salaries of working professionals and football players- where the latter were compensated with almost six times of the former, sky rocketed (Frick, 2008). The disagreement thus compelled the Germany football associations to cap the transfer prices to a maximum of 50,000 Deutsche Mark (DM) and the set the range of players’ salaries from 250 DM to 1200 DM per month. But with time, football fans' apparent lack of concern for the level and progression of player salary was hardly surprising given the consistently rising earnings from ticket sales and merchandise. However, this development, which can be largely linked to the growth of the TV earnings produced by the clubs, has most recently caught the attention of a number of politicians (Frick, 2008). Frick (2008) demonstrated that the amount of money earned each year was directly correlated with the number of goals and games played the previous season. Another intriguing finding on the significance of the player's birthplace was also established.

Therefore, chapter 3 of the research aims at providing an encompassed study of the wages of the players from various different football leagues across Europe.

**2.3 CLUB REVENUE**

Football clubs' three primary income sources historically have been match money such as ticket sales, media and broadcasting income as well as commercial income from sponsorship and merchandise. By concentrating on the identification, growth, and ensuing profitable on-sale of players, some clubs have created a unique business model. Transfer fees have become into a crucial source of cash for some of these clubs. The clubs in the EPL are rather big spenders, acquiring seasoned players from other top leagues in Europe and beyond. The most significant financial consideration for clubs in contemporary professional football is player pay, along with transfer fees. Szymanski and Smith (1997) have demonstrated a statistically substantial and favourable correlation between teams' personnel expenses, such as player salaries and transfer costs, and their athletic success.

The level of competitiveness in the football transfer market is frequently used as a crucial starting point in current studies (Carmichael, 2006). On one extreme, there is the idea that the football transfer market is monopolised by clubs and players (Carmichael, 2006), with a single team seeking to sign a certain player. According to this perspective, clubs and players bargain about transfer payments (Carmichael and Thomas, 1993). Because of the asymmetric knowledge about a player's quality and commitment, the transfer market is characterised by ambiguity, and there is a risk because it is unknown, prior to a transfer, how well a player would perform in the new team (Carmichael, 2006). On the opposite end of the spectrum, another viewpoint contends that the transfer market is competitive and that transfers fees are set in a competitive procedure which has been possible post the Bosman ruling (Carmichael, 2006). Contrary to the situation in the monopoly market, the transfer market is characterised by contract freedom, significant possibility for mobility, a large number of buyers and sellers, and the availability of detailed information on players' performances (Carmichael, 1999).

Since a deal is established between the buying and selling clubs in monopoly markets, the usual *independent variables used are the inherent traits of the buying and selling clubs, such as team performance and market size* (Vrooman, 1996). The monopoly market is distinctive due to this quality. The *ability and human capital of players are typically used as independent variables for the competitive model. These variables include age, experience (number of league appearances), goal record, position played, international appearances, selling club status and performances, divisional standing, and goal total* (Vrooman, 1996)*.* The process of determining a player's transfer fee in football in a market that is competitive depends on the player's inherent talent and human capital, which are primarily reflected in their marginal revenue product at the selling club. According to this hypothesis, players not only desire to join more successful teams, but those teams also want to be identified with the former, and this influences transfer fees significantly (Szymanski and Smith, 1997).

Therefore, it becomes very important to find the extent up to which the wage bills of the players and the club revenues are correlated. A deeper dive into the issue will also help in shedding light upon the fact how a handful of clubs dominate the player market with respect to acquiring the best talents. Chapter 6 deals with the machine learning model which explains how the wages of the players are impacted by club revenue.

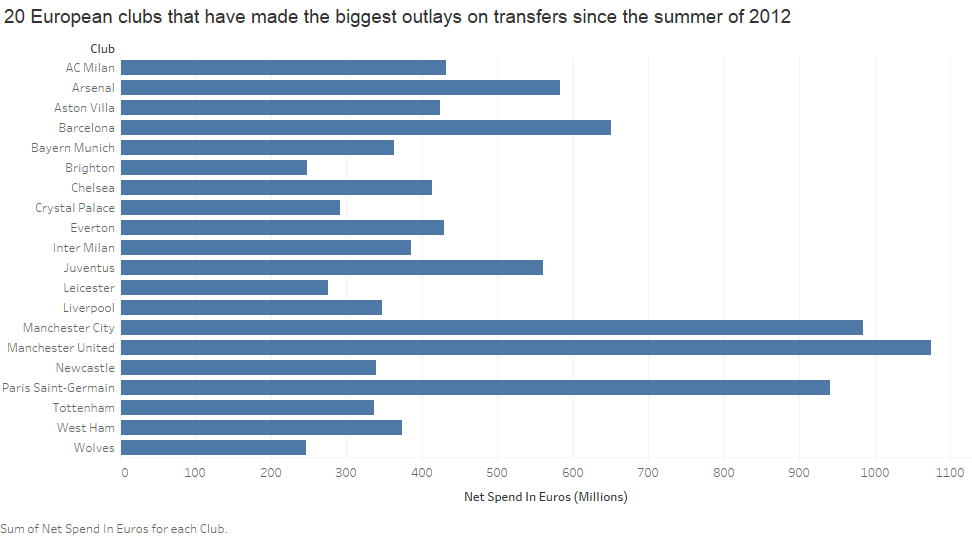
**2.4 FURTHER ANALYSES**

The studies and comments provided from the perspective of following analyses in the dissertation are purely theoretical. The following will aid in accounting for the gaps or bolstering the claims based on the facts already established in the previous studies.

**2.4.1 TRANSFER MARKET**

The term "net transfer expenses" refers to the total of transfer fees collected and the cost of acquiring new players. In the current market, transfer fees are an example of discretionary spending, as opposed to wages, which are required by an existing contract. Football player transfer prices have grown to be a big industry concern, and as a result, there is a need for an explanation of the factors that affect transfer pricing (Ezzedine, 2020). Economists have worked on various econometric models to assess and explain the transfer prices using both monopoly and competitive markets, resulting in interesting findings and establishing certain facts. The specification of the econometric model and the independent variables used to determine the transfer prices depends highly on the type of market in question.

Ezzeddine, M (2020) explained in detail how football transfer pricing was conducted in real-time using econometric approach. The first body of literature placed a strong emphasis on calculating football transfer fees, the second set on valuing football transfers, and the third set on using machine learning techniques to comprehend football pricing and football in stock market. *The study established that the transfer fees and salaries have soared high and the related trends are incomprehensible*. This very well supports the proposed hypothesis discussed in the subsequent research here. With big clubs garnering all the sponsorships and broadcasting rights, the bridge between the revenues of rich and poor clubs have widened. Goal-scoring record, league appearances, international appearances, and age (and occasionally age squared) are among the factors that are typically shown to be statistically relevant in influencing football transfer fees (Carmichael et al., 1999).



Graph2.2 European clubs 20 highest transfer expenditure since 2012

*Football players' transfer fees are significantly influenced by their age (up to 26 years old) and the amount of games they have played* (Ruijg and Ophem, 2014). According to Speight and Thomas (1997) and Carmichael et al. (1999), factors including a player's age, the number of games they have played for a club, and their number of international caps have a substantial impact on both their salary and transfer fees (games with the national team). Since Szymanski and Kuypers (1999) expanded the discussion on transfer fees by proving two key hypotheses, which were well known to viewers and supporters of English football, more universal results have been discovered. They first demonstrated that clubs who invest more in their players had more on-field performance, and they then demonstrated that a club's profitability is based on league place. Szymanski (2015) also demonstrated how the transfer system keeps top clubs dominant by ensuring that they are the only ones with the financial wherewithal to pay the high transfer fees required for the best players. The way the system is set up right now not only treats players unfairly, but it also encourages behaviour that was never intended.

Ezzedine, M (2020) has also explained in detail the problem of selection bias and has opted for James Heckman’s theory to resolve it. Although a fairly detailed study with outcomes concluded that still stand, the study of stock market changes post covid have not been considered. Additionally, current papers have no explanation provided for the specifics of the superstar impact in football or the identification of reliable market categories for transfer pricing. Only consistent data and methodologies may be used to address this. They suffer from a selection bias that they are unable to measure, despite the fact that Carmichael et al. identified it as early as 1999.

**2.4.2 MAJOR FOOTBALL EFFECTS**

With the analyses on the wages and the conclusions drawn from it, two major football effects would also be briefly used to comment on the results towards the end namely, the superstar effect and the tournament effect. Football superstars are known to earn far more money than average players, and the scarcity of these players makes the effect stronger. Superstar football players draw monopsony rents, as demonstrated by Garcia-del Barrio and Pujol (2007). This is because there are many clubs vying for their services, causing clubs to raise salaries to remain competitive and sign these players. Other arguments have been made to show that, despite the fact that a player's performance increases with absolute income, wage inequality can have negative impacts on other players and have an impact on coaching choices (Torgler, et al., 2006; Garcia-del Barrio and Pujol, 2007). For the latter, it has been seen that, in contrast to other players who are paid less generously, coaches frequently employ players with greater incomes in ways that are out of proportion to their on-field performance.

**Superstar effects** occur when a player receives a larger wage than his or her teammates as a result of increasing the club's revenue.

**Tournament effects** are when wage discrepancies are based on relative differences between the people rather than marginal output.

Since the days of classical and neoclassical theory, superstar economics has developed, with its study becoming considerably more evident and pertinent in present times. Two of the highly popular but conflicting theories that have been floating around are those given by Rosen and Adler where Rosen in 1981 put out an idea that tiny disparities in talent can lead to significant inequalities in pay in markets where productivity and revenue have a convex relationship while Adler in his 1985’s study asserted that popularity disparities rather than differences in competence may be the cause of superstar pay.

Sherwin Rosen's study from 1981 is the first example of superstar economics in contemporary literature. According to Rosen, the "superstar phenomenon" is when "very few people earn tremendous sums of money and rule the activities they engage in". He bases his theory of superstars on their talent. The distribution of skill and the distribution of rewards have a convex connection, according to him, WHICH leads to superstars when "little variations in talent become amplified in big earnings discrepancies." In order to explain the wealth of superstars, Rosen additionally introduces the concepts of imperfect substitution (goods with a lower level of substitutability) and joint consumption technology (sellers can service the entire market without the costs of production increasing). Customers are willing to pay more for superstars because they perceive marginally less brilliant performers as drastically inferior alternatives. Additionally, because production services can be duplicated at a constant marginal cost, superior performers can readily extend their impact more frequently, maintaining their market dominance and raising their superstar salaries. Rosen's theory is heavily founded on the idea that in a market where performance is valued, a superstar's talent, even if it is only marginally superior to that of its rivals, results in significant earnings differences.

Moshe Adler's 1985 paper is the second piece of literature on superstar economics. Adler approaches the superstar phenomena from a different angle. He contends that in order to appreciate an artist or athlete's or athlete's performance, people must learn more about them through interaction with others. Superstars consequently emerge when more people become aware of a certain individual, even if they just possess comparable or marginally superior talent. A superstar is created by a snowball effect and accepted by the public to minimise their learning costs, according to Adler, who thinks that a star's knowledge is built up through a positive network externality (the quantity of a good is demanded by a consumer increases in response to an increase in purchases by other consumers).

According to Adler's theory, the media has a key role in the production of superstars since they have an impact on which performances garner public attention. Adler's argument is mostly grounded in the popularity of superstars and how, despite having comparable talent to rivals, their popularity within a performance market promotes higher earnings.

By combining real-world data with information gathered from the football video game FIFA 2015, Shin and Gasparyan (2014) conducted research on forecasting football results. They combined real data with player characteristic data from FIFA 2015, such as heading, passing, shooting, and strength, to forecast match outcomes, and the results are better than when they used only real data alone. Additionally, they claimed that collecting data from video games might save a lot of time and effort because it can be highly expensive to calculate or get some data from the actual world.

The components in football market encourage economic analyses especially in terms of labour market. Although the comparisons and similarities drawn by the author are purely from that perspective yet it highlights the controversial issue of income inequality which from a behavioural perspective does have a negative impact on contemporary players. Recent agreements limiting compensation increases for star athletes were reached by team owners in the major North American sports. Maybe it's just a coincidence that league efforts to control salaries have occurred at a time when both the popularity of sport and player salaries has increased significantly. Individual players' pays are limited by salary limitations. Payroll restrictions place a cap on the overall amount of wages that each team can spend, although they do not directly affect how much a player is paid. But similar wage caps have not been implemented on teams under FIFA although there have been claims about informal caps. This presents a need for the wages to be studied such that players’ real worth from the perspective of wages can be analysed and other “non-superstar” players can also have their worth assessed and paid reasonably.

**2.5 CONCLUSION**

Many studies have been explicitly conducted on the transfer prices market. There has been a theory constantly popping in most of the papers where transfer prices are directly proportional to the wage bill of the clubs. This ultimately leads to the clubs over-paying their aging but star players just to maintain their league status. Pedro Garcia-del-Barrio considers the soccer industry to be a dual-labour market.

Due to very limited studies in English, this approach can be used as a supporting technology for compensation negotiations between the agents of the football players and clubs. Additionally, it can be used for conducting quantitative analyses of relationships between pay and performance. The method takes into account aspects that are not related directly to the game such as the popularity of the player among fans, predicted merchandise sales, etc., which could have a significant impact on pay, particularly for elite players. The talents that most influence prices, according to an analysis of player earnings in five European football leagues, are basically the same across leagues, but there are some differences. Analysis of underpaid and overpaid players reveals that underpaid football players have superior responses, vision, acceleration, agility, and balance while overpaid players tend to be stronger. The idea of this study is to present an all-around study of the prediction of football wages. The studies are based on the analysis of player salaries in major European football leagues; the report also aims to determine if the parameters that are highly likely to affect the salary are consistent across leagues. Therefore, a productivity based study will be mapped with popularity based explanation

CHAPTER III  
DATA

**3.1 PART A- DATA FOR WAGE PREDICTION MODEL**

The data used in the experiment of football players obtained from sofifa.com, which offered details on each player's performance and abilities as well as their pay. The football club's gross salaries, as agreed upon by the club and the player's agency, are the wages considered in this study. They do not include the player's other potential sources of revenue, such as endorsement deals or sales of goods. Football teams are obligated to declare the right salary in order to comply with rules such as the Financial Fair Play due to the severe rules of football organisations like FIFA and UEFA, and because these rules are strictly enforced, the reported pay can be trusted. The dataset initially had 16710 rows and 54 columns.

Although past experience has clearly shown some instances in which these rules have been broken in an effort to conceal financial information, these instances can be viewed as exceptions, and it can be expected that the salaries declared by football clubs are generally reliable. Using a web crawler, which gathered each player's data during the 2021-22 season, the information from sofifa.com was obtained. Football experts initially assembled the dataset, analysing each player's skills for the goal of recreating them in video games1, and subsequently for scouting2 (Yaldo and Shamir, 2017). These statistics have proven to be comparable to or superior to other sources of football data and have been successfully used for predicting the results of football matches ( Shin and Gasparyan, 2014). The measured skills are expressed as integers between [0,99]. The player salaries are also included in the dataset, which is useful for scouting purposes. Table3.1 is shown below representing variables from the dataset and description

**Table3.1 Wage Prediction Dataset description**



Two more columns, ‘league’ and ‘country’ of the league were taken from sofifa.com and added to the dataset. Attributes such as logo, jersey, photo, etc. were also removed since they were in no way going to contribute to the model.

**3.2 PART B- DATA FOR CLUB REVENUE ANALYSIS**

Furthermore, the data for the club revenues and its composition such as matchday, broadcasting and commercial, etc. have been taken from the financial reports published by the respective clubs annually. The data for transfer prices and market values including transfer income and transfer expenditure have been taken from tranfermarkt.co.uk and capology.com. Deloitte’s annual financial report was also used to verify the data of the top 10 clubs. The variables used have been given below with their description-

Table 3.2 CLUB REVENUE ANALYSIS DATASET



CHAPTER IV   
METHODOLOGY

With ‘wage’ as our continuous variable, regression model was used to predict the wages of the football players. The wages have been predicted in two ways. The first part explains the wages from the perspective of players’ performances, market values and skills. The second part explains the wages from the perspective of the club revenues in order to understand how different types of clubs garner their income and compensate their players.

**4.1 ASSUMPTIONS AND HYPOTHESES**

The professional football market considered across the analysis was ‘Competitive market’ where players’ individual performance and skill attributes were considered for the analysis.

This paper aims at determining how performance affects the wages and other aspects that in return affect a football player's transfer value aside from their performance. The variation in talent and performance are the main causes of the discernible variation in player salaries:

A player's salary will rise with success (league appearances, goals), age (experience), performance (overall rating) and popularity (appearances in the national team). Player compensation will also be greatly impacted by the clubs' various financial capacities, which in turn depend on the size of the relevant markets, the club's history, and its athletic accomplishments.

**4.2 WAGE PREDICTION BASED ON PLAYER PERFORMANCE AND MARKET VALUE**

**4.2.1 DATA ANALYSIS**

The study was conducted using supervised machine learning algorithms where a target variable was identified and then the final model was created through modeling relationship between the target and predictors. The target variable was then predicted using **regression** techniques. Based on this relationship, new values for the target output were predicted. Several supervised learning techniques such as Random forest, decision tree, regression model and neural networks were deployed to yield the outputs. The wages of 16710 football players across various football leagues formed the target variable with the players’ performances, skill, brand value and ratings constituted the predictors. The RMSE of these models were compared in Chapter 4 to identify the best model that can be used to validate the objective of our study.

Following columns were omitted from the dataset as they were not deemed useful in contributing to the analysis in any way.

Table 4.1 Variables removed



**4.2.2 DATA PRE-PROCESSING**

Data preprocessing is a crucial phase in the machine learning process since the quality of the data and the information that can be extracted from it directly influence how well our model can learn. For this reason, it is crucial that we treat the data before introducing it to the model (Kumar, 2018). Within the scope of pre-processing, concepts such as handling null values, treating duplicate values, etc. were evaluated and used accordingly to deal with our dataset.

**4.2.2.1 NULL VALUES**

When the dataset does not have data stored for specific variables or participants, we have missing data, also known as null values (Saha, 2019). Data loss can occur for a variety of reasons; including incorrect data entry, device failures, lost files, and more. Missing data can occasionally result in sample bias, depending on their type, which makes them troublesome. As a result, the findings might not be extrapolated from the study because the data would have been collected from an unrepresentative sample.

Using Python, the data was vetted for null values in order to ensure a high-quality and less redundant dataset for analysis. The number of null values found has been given below-

Table 4.2 Null Values



In order to treat the null values and impute them, the presence of outliers was detected first for continuous variables in order to finalize the type of descriptive statistic measure to use.

**A. Categorical Values**

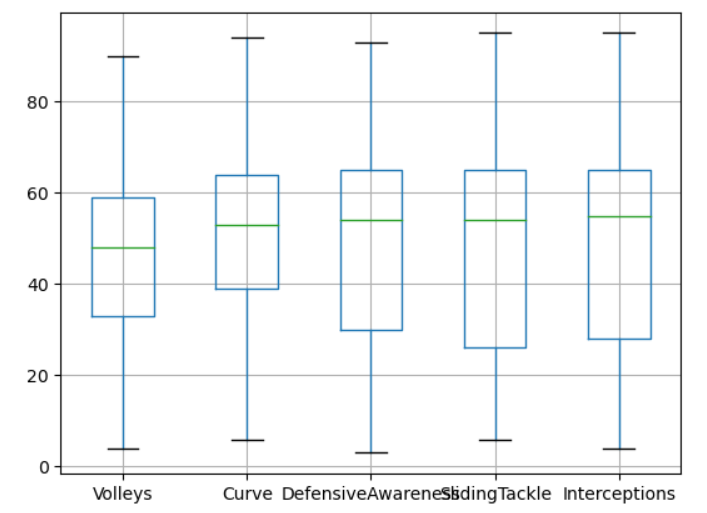
Being categorical variables, Club, Position and Contract Valid Until were substituted with “Mode” values in order to account for the null values.

**4.2.2.2 OUTLIERS**

Values that stand out from the majority of other data points in a dataset are known as outliers, and they can have a significant impact on the statistical analysis and skew the results of any hypothesis testing (Pritha, 2021). Box plots were used to check the presence of outliers.

**B. No Outliers**

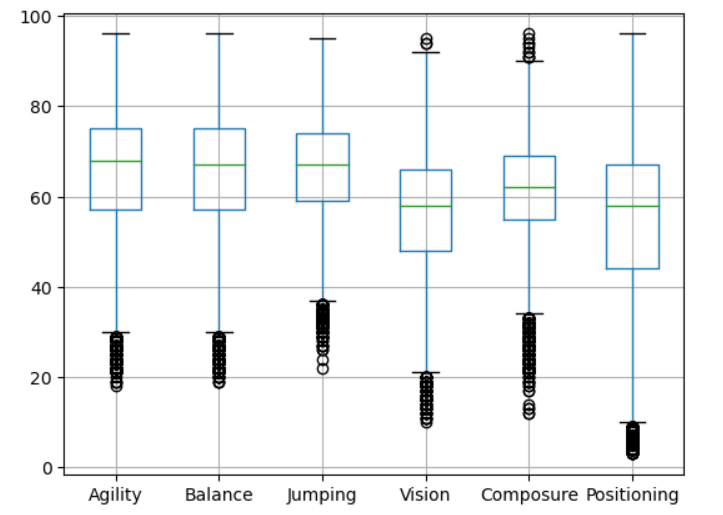
Continuous variables such as Volley, curve, DefensiveAwareness, SlidingTackle, Interception did not have any outliers and therefore the null values for these features were populated with their respective mean values.



Graph 4.1 Variables with no outliers

**C. With Outliers**

Furthermore, other continuous variables such as Agility, Balance, Jumping, Vision, Composure, and Positioning had outliers present and therefore, the null values in these attributes were populated with median values.



Graph 4.2 Variables with Outliers

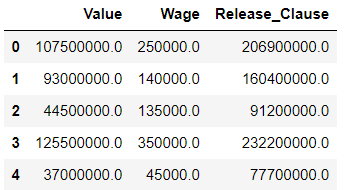
There are various ways of treating outliers such as substituting the outliers with mean, median, capping or deleting them altogether. But with this real time data, the most logical solution was to not treat the outliers as the extreme high and low values were going to acknowledge the vast differences among the salaries and accordingly predict the wages.

**4.2.3 EXPLORATORY DATA ANALYSIS**

Data scientists use exploratory data analysis( EDA), which uses data visualisation ways, to examine and assay data sets and summarise their worthwhile parcels. It enables data scientists to find patterns, identify anomalies, test suppositions, or corroborate hypotheticals by determining how to modify data sources to achieve the answers they need (IBM Cloud Education, 2020). Comprehending the data and trying to extract as numerous perceptivity from it as possible is a smart strategy. Before getting into the data with analysis, EDA focuses on making sense of the information formerly available.

EDA's major thing is to encourage data analysis before making any hypotheticals. It can help in chancing striking crimes, better understanding data patterns, spotting outliers or unusual circumstances, and discovering interesting connections between the variables. To make sure the findings they produce are dependable and applicable to any asked business objects and pretensions, data scientists can employ exploratory analysis (IBM Cloud Education, 2020). EDA aids stakeholders by assuring them that they're posing the proper questions.

Table 4.3 Top 5 values after removing text

To begin with, certain minor tweaks were done to the dataset such as renaming certain columns in order to make the dataset suitable to have exploratory data analysis performed on them. For instance, the values in columns ‘Value', 'Wage' and 'Release\_Clause' had texts ‘M’ and ‘K’ concatenated to the numbers making it hard to represent graphically. Therefore, the texts were removed and the data were converted from millions and thousands to numbers:

These columns were again vetted for outliers and then treated accordingly. Statistical summary for these columns were then drawn out.

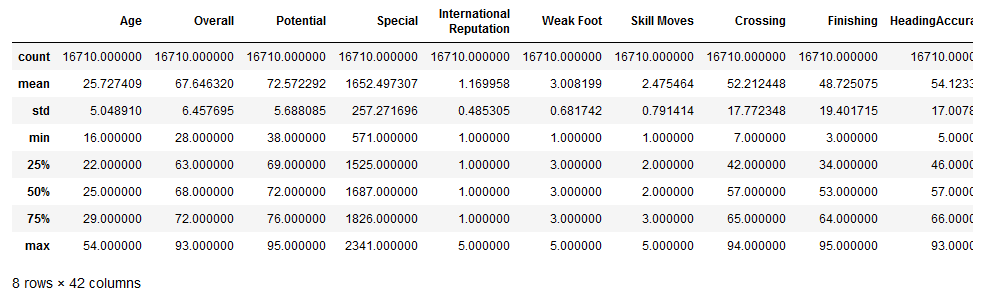
|  |  |
| --- | --- |
| Graph 4 3 To check outliers and anomalies | Table 4.4 Descriptive Statistics |

**4.2.3.1 DESCRIPTIVE STATISTICS**

The parcels of a data set are organised and summarised using descriptive statistics. A data set is a compendium of compliances or responses from a sample of a population or the complete population.

In quantitative exploration, describing parcels of the responses, similar as the normal of one variable or the relationship between two variables, comes first in the statistical analysis process after data collection.

Table 4.5 Descriptive Statistics of all the variables



Distribution, central tendency, and dissipation are the three primary orders of statistics that are described.

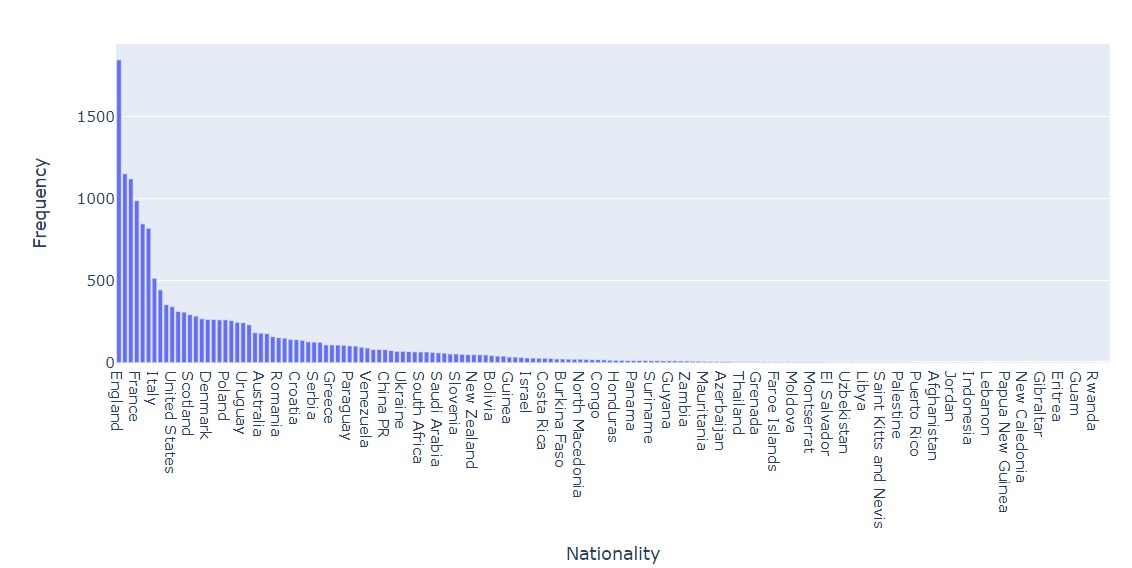
The distribution relates to how constantly each value occurs. The primary trend is related to the value pars. The dissipation or variability refers to how unevenly distributed the results are.

For all the available continuous variables, central tendency measures were explained. For example, on an average the “Overall” rating is 67.6 with 28 being the minimum and 93 being the maximum ratings.

**4.2.3.2 GRAPHICAL REPRESENTATIONS**

**1. COUNTS OF PLAYER FROM COUNTRY**

Football has always reflected an essence of diversity when it comes to its professional players. Majority of the players hail from England as per our dataset followed by France and Italy only a little behind it.



Graph 4 4. COUNTS OF PLAYER FROM COUNTRY

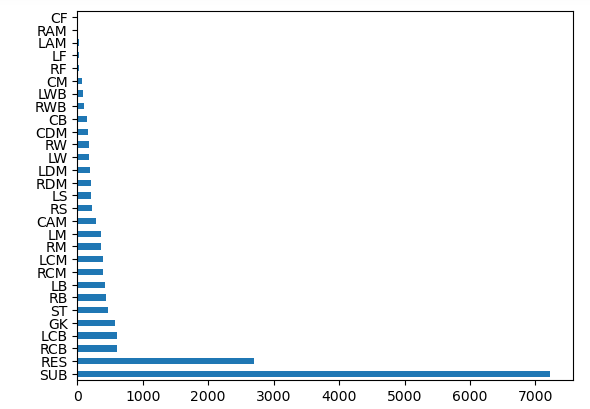
Table 4.6. COUNTS OF PLAYER FROM COUNTRY



**2. COUNTS OF PLAYER PER POSITION**

The graph is based on the distribution of each player's position during the game captured by the column “Position”. This field differentiates well between the active players and the ones that are substitute or reserved. Therefore, setting the substitutes and reserves aside, Right Center back is the most prevalent position with 611 players playing from that position, followed by 609 left centre back and 572 goalkeepers.

Table 4.7 Counts Of Player Per Position

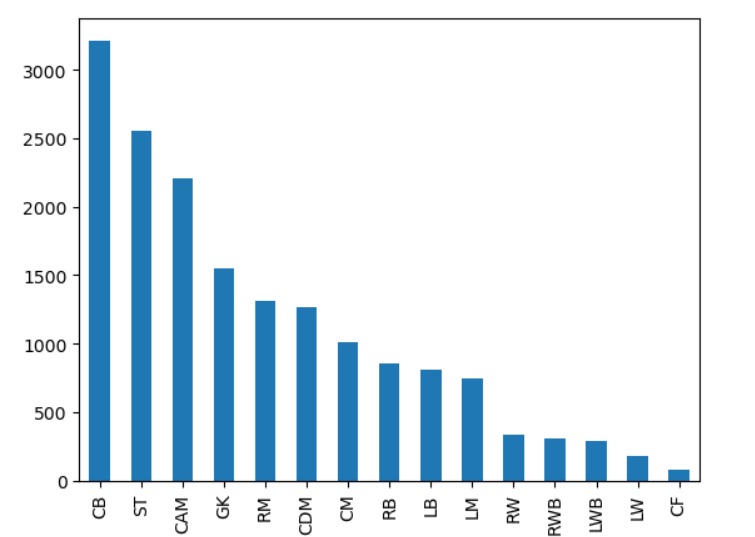


Graph 4.5 Position vs Frequency (Counts of Player per Position)

**3. BEST POSITION DISTRIBUTION OF PLAYERS**

**A. BEFORE REGROUPING VALUES**

The bar graph shows that majority of the players which accounts for almost one-third of the player pool that play at Center Back position followed by strikers and Center Attacking Midfielder. There is a scarcity of players who play as Center forward preceded by Left Wing.

.  
****

Graph 4.6 Best Position Values- Before regrouping

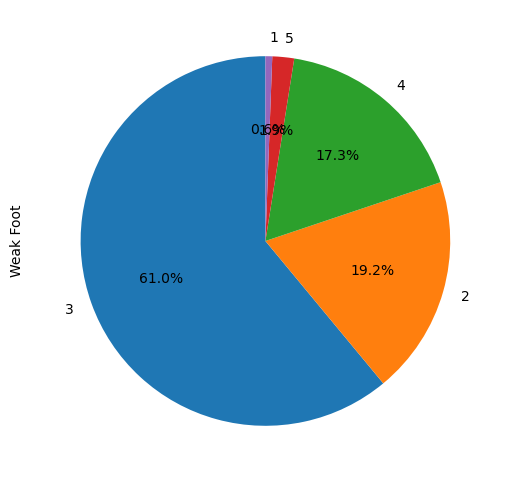
**B. AFTER REGROUPING VALUES**

To simplify the study further, the data points of the column “Best\_Position” were regrouped into four basic positions viz. defender, attacker, goalkeeper and striker. According to the plot, majority of the players play in the midfield position.



Graph 4.7 Best Position Values- After regrouping

**4. COUNTS OF PLAYER PREFERRED FOOT**

****Majority of the players prefer playing with their right foot. This calls for formulation of strategies for practice and skill development within clubs where there is a need to deploy cross-combination of footedness and wing position to improve gameplay. For example Jaden Sancho, a left-footed Manchester United player plays from Right wing while Antony, a right-footed player plays from the left wing. But due to less number of left footed players, it is difficult to find a player especially the one who specializes in the wing position. This leads to higher market values for such players as evident from where Jaden Sancho ranks number 9 with respect to his market value.

Graph 4.8 Counts of Player Preferred Foot

**5. PLAYER COUNT FOR WEEK FOOT**

A player who has a 5 star rating for weak foot has equal performance precision, power, and tackling ability with both feet. 17.3% of players have an additional 20% probability of error with a 4 star weak foot. Almost 61% of the football players have a 3 star weak foot which implies that there is a 40% chance of error.

Graph 4.9 Player Count for Week Foot

**6. SKILL MOVES**

**A. SKILL MOVES VS PLAYER COUNT**

The pie chart depicts that only a mere 0.4% of the players i.e., almost 66 players have the most efficient skill set while 39.4% players or 6583 players have average skill moves and 9.5% or 1587 players with least rated skill set.

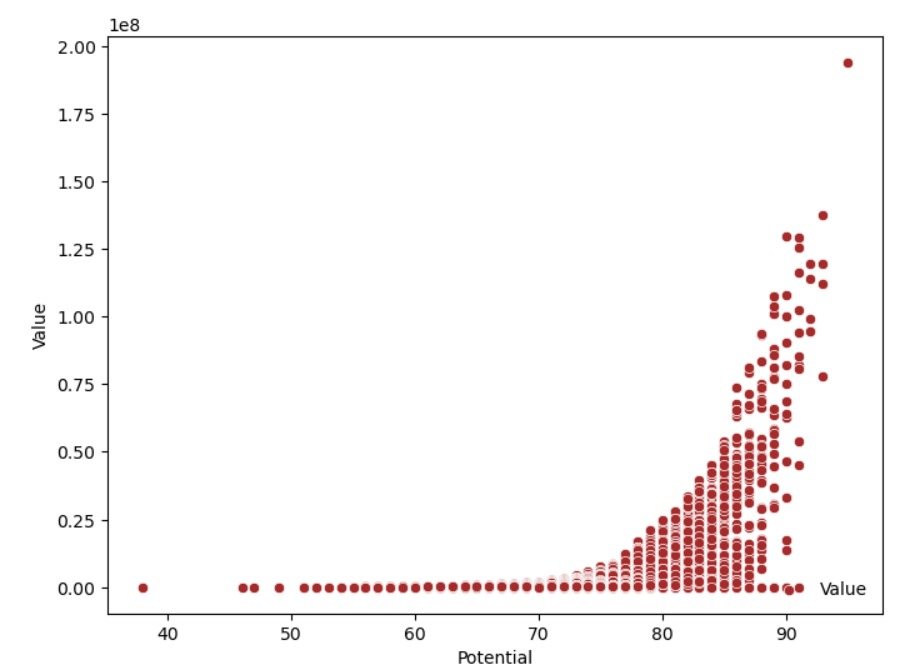
**B. SKILL MOVES VS WAGE**

When mapped against wage, the bar graph on the right clearly shows that the highest compensation is for the highly rated skill set which is 5. Despite having small concentration of players for the said skill rating, the players account for a substantial of amount salaries being paid.

|  |  |
| --- | --- |
| Graph 4.10 Skill Moves vs Player Count | Graph 4.11 Skill Moves vs Wage |

**7. POTENTIAL VS VALUE**

Potential is concerned with a player's potential future prowess.

****

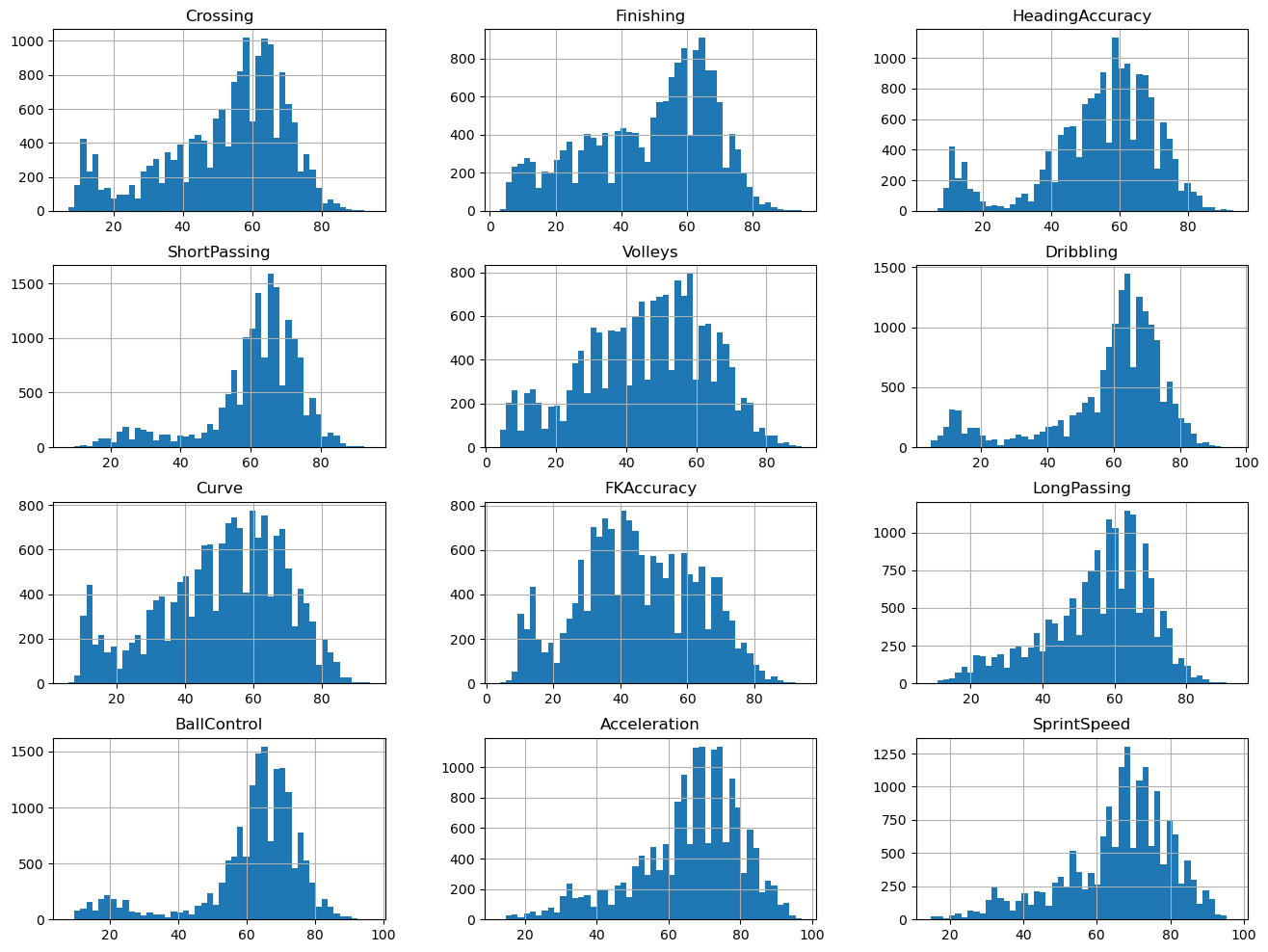
Graph 4.12 Potential vs Value

Having joined the club in his 20s, the player has a potential of less than 80.   
If the player's potential is between 80 and 84, it implies that the player is "showing outstanding promise".   
A "fantastic prospect," is for the players in the potential range of 85-89.   
The player’s potential of 90-94 gives him a high probability of being special.

While Value column gives an estimate of a player’s market price in Euros. Therefore, the graph shows the relationship between the potential of a player to excel in the future and his current value based on which a club can acquire the player and groom him accordingly for the benefit of the club.

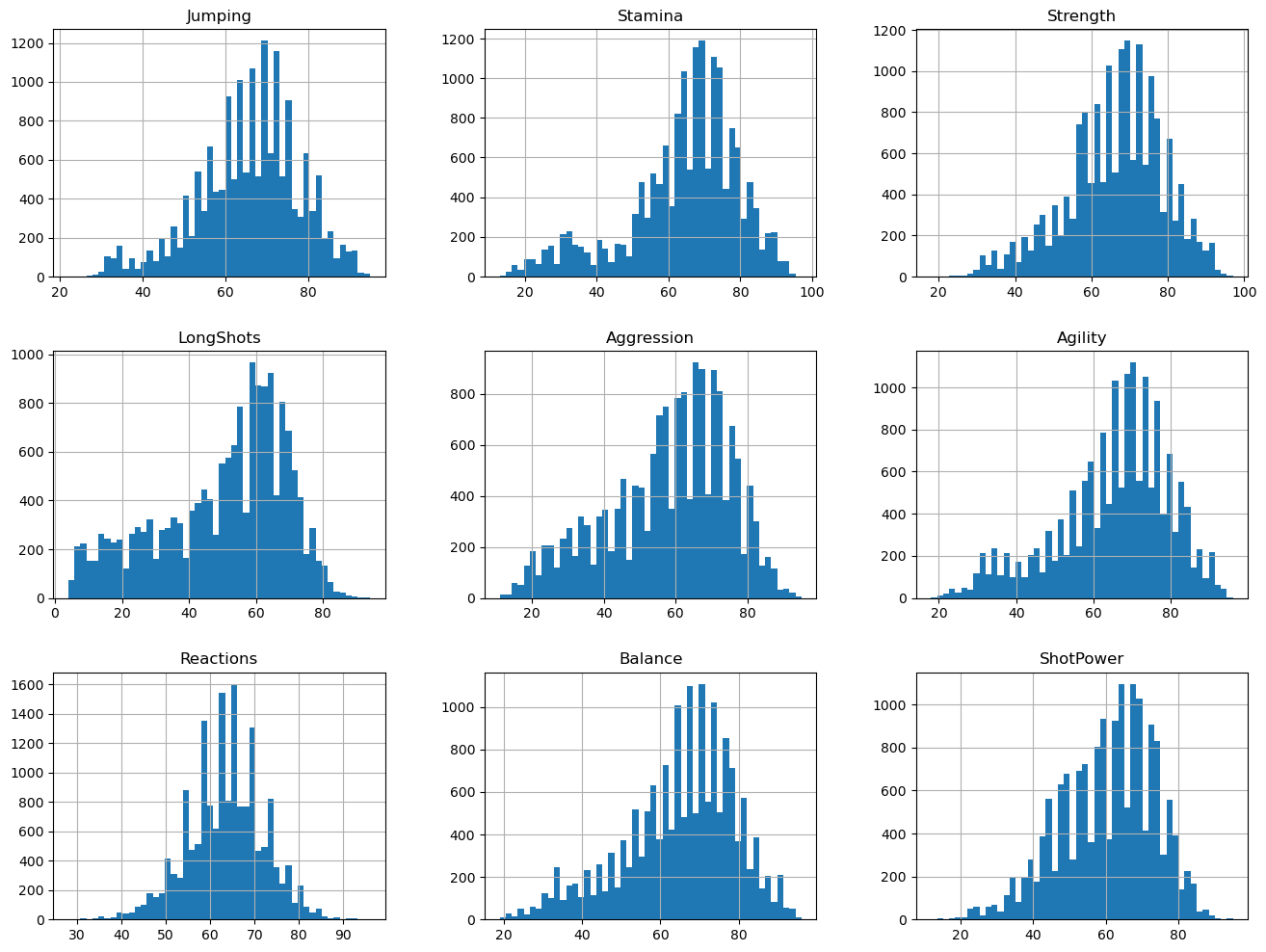
An increasing trend can be seen for the players with potential 74 and onwards.

**8. DISTRIBUTION OF PLAYERS’ SKILL RELATED ATTRIBUTES**

****

Graph 4.13 Frequency Distribution 1

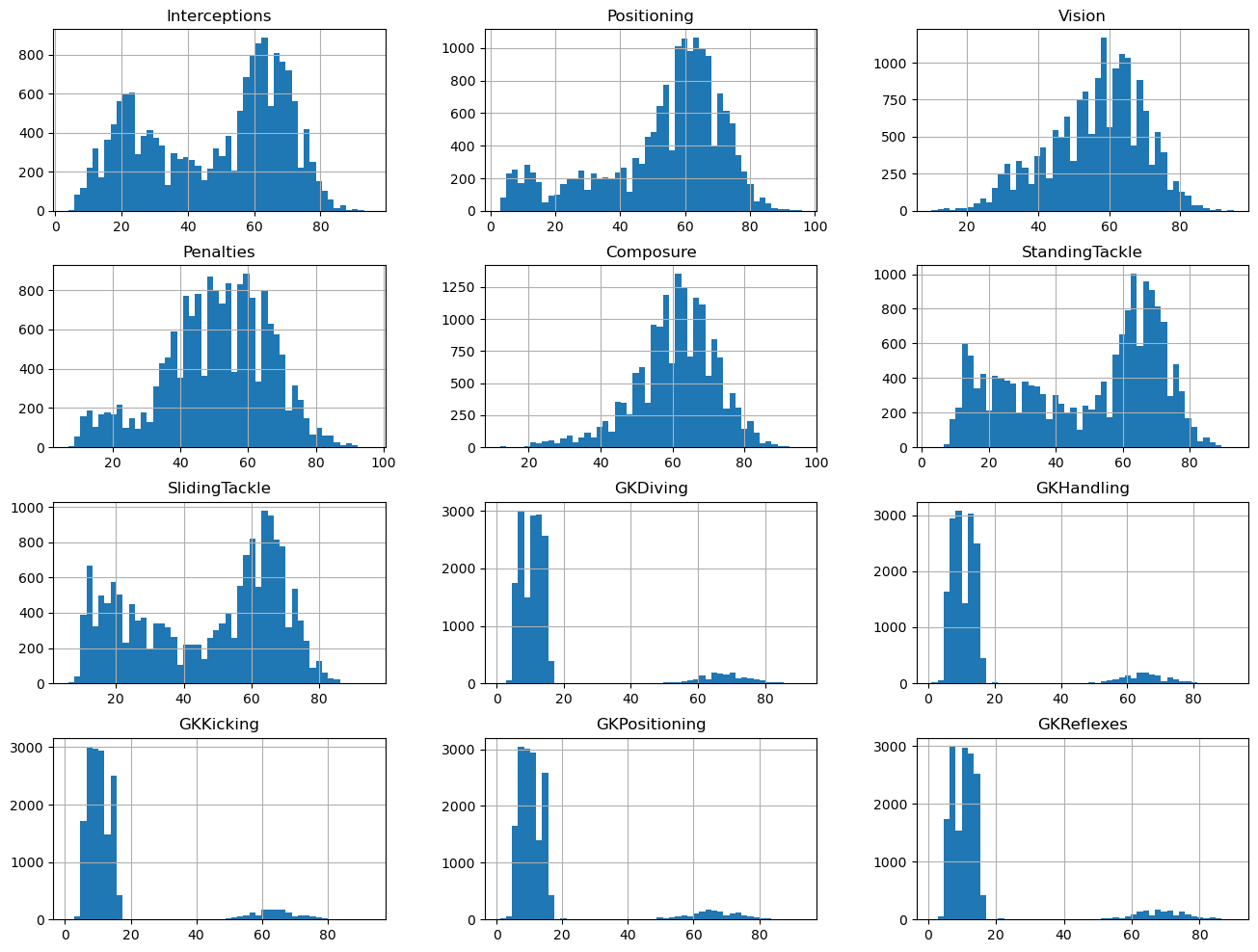
This section demonstrates the frequency distribution of all the skill related attributes. The distribution of Variable Reactions is roughly symmetric. It is interesting to see that the ball control, stamina and interception histograms show bimodal distributions. Goalkeepers are the reason for the smaller modes.

****

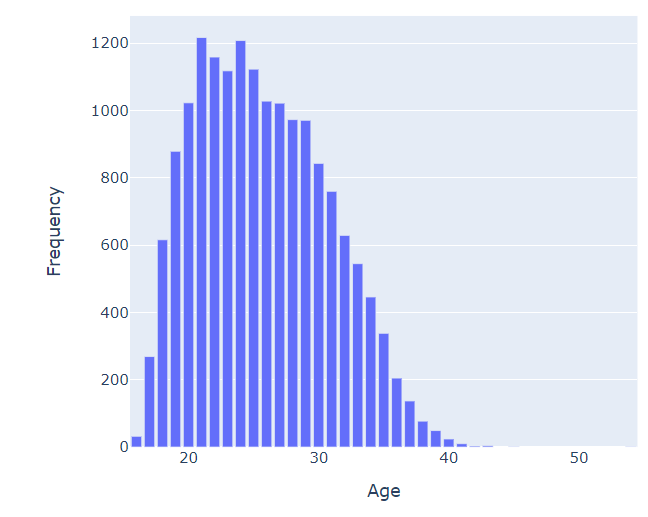
Graph 4.14 Frequency Distribution 2

In addition to this, Balance, agility, composure and vision have symmetric distributions while SlidingTackle is a bimodal distribution.

The goalkeepers’ attributes such as diving, handling, kicking, positioning and reflexes with prefix “Gk” are highly right skewed. This implies that there are less number of players with enhanced abilities in all these aspects as compared to the others.

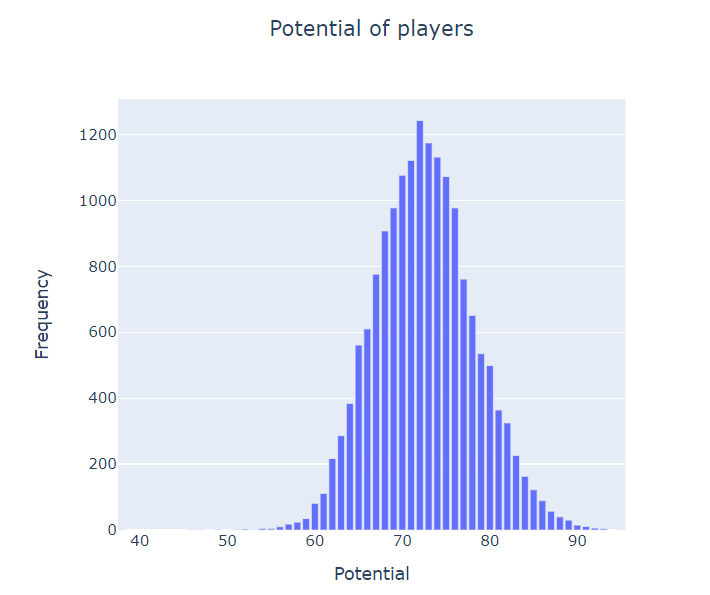
****

Graph 4.15 Frequency Distribution 3

**9. AGE DISTRIBUTION OF PLAYERS**

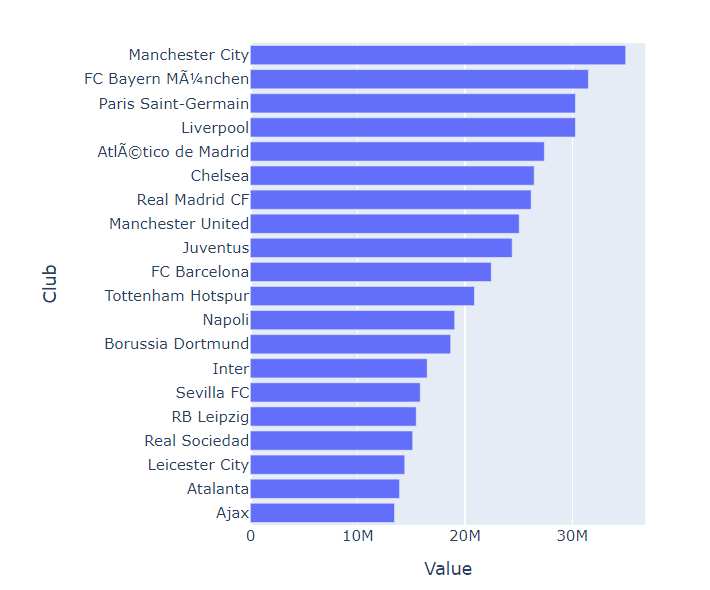
Graph 4.16 Age Distribution Of Players

Age plays a significant role in the world of sports as the older the player, the less they tend to be picked by a club due to factors such as reduced stamina and activity. The majority of the players lie in the age bracket of 18 and 33. The number of players with growing age decreases with only a handful of active players older than 40.

**10. POTENTIAL OF PLAYERS DISTRIBUTION**

Graph 4.17 Potential Of Players Distribution

The graph shows the potential ratings of each player rated from 0-99. There are more than 1200 players with a rating of 72 while the range of the potential ratings of players lie somewhere between 54 and 94.

**11. CLUBS VS PLAYER'S AVERAGE VALUE**

Graph 4.18 Clubs vs Player's Average Value

The graph below shows the market value of players in millions averaged across the clubs, against the clubs that own them. The graph represents 20 European clubs who have the possession of the most of the players with highest. With Manchester City at the top, English Premiere League holds the majority with 30%.

# 12. PLAYER BASED ON VALUE VS OVERALL

Based on the overall rating, the top 15 players have been tabulated in a decreasing order. Topped by a striker, Kylian Mbappe is one of the best players around. The English Premier League players dominate the list with almost 50% of the players.

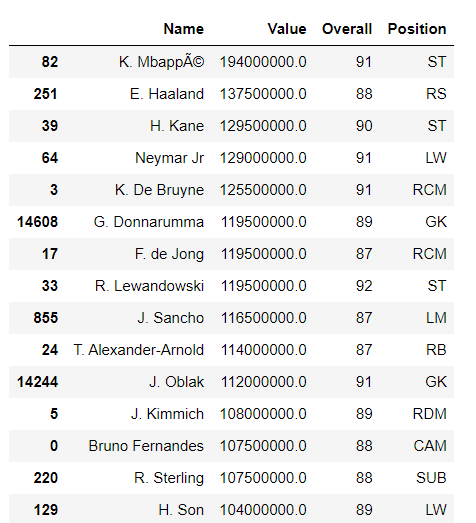
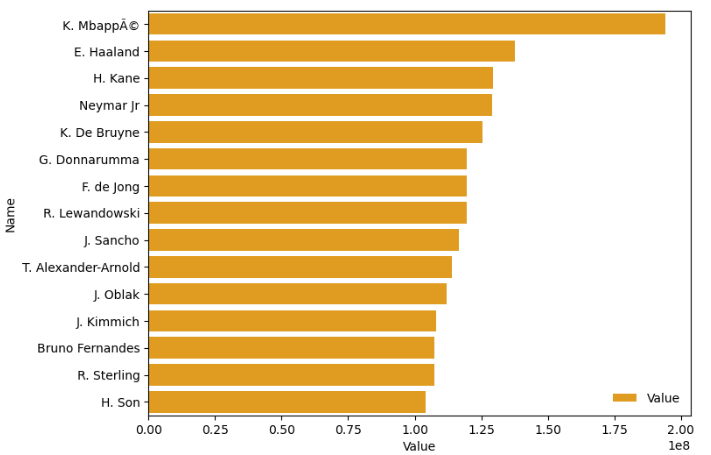


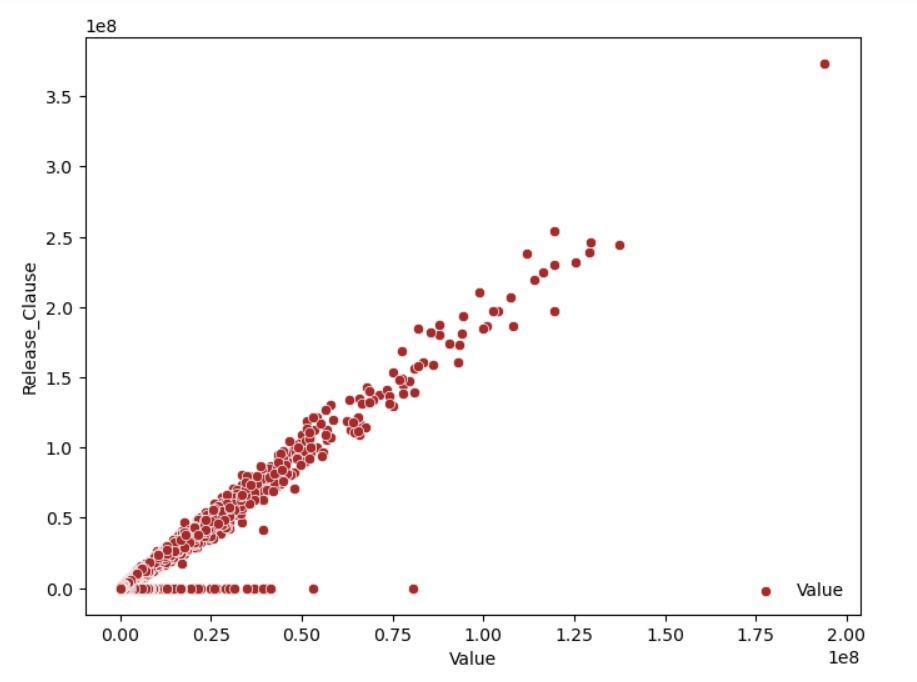
Table 4.8 Player Based On Value vs Overall and Position



Graph 4.19 Player Based On Value

**13. RELEASE CLAUSE VS VALUE**

The graph shows an almost proportional increase between the market value of players and release clause values. A highly valued player will have a higher fee as per the release clause amount which implies that in order for a club to buy a highly valued player from a team before the contract expires, the buying club will have to pay the selling club the amount of the release clause. Since release clauses are always so high, the smaller clubs cannot afford to buy the talents and mostly rely on building players picked from their academies. The release clause values for every player is also higher than the market value for that player.

****

Graph 4.20 Release Clause vs Value

**14. TOP 15 RELEASE CLAUSES**

The table above shows the top 15 players with the highest release clauses being graphically represented below as well. Kylian Mbappe, with the highest market value as evident from the previous graph, again tops this list. In fact, 13 players from the Table are also in present Table showing a high collinearity between the columns Value and Release Clause. The English Premier League players again dominate the list which somehow accounts for the fact how the top teams such as Manchester City holds a monopoly in the “competitive transfer market” just because of deeper pockets than the rest of the clubs.

|  |  |
| --- | --- |
| Table 4.9 Top 15 Release Clauses | Graph 4.21 Name vs Release Clause |

**15. AGE VS VALUE**

The market value of a player depends on his age as players between the ages 22 and 33 have been valued exceptionally high. This is mostly due to the reason that a young player is expected to be more active and healthy compared to older players (above the age 35). Even here, the relationship of value with age is not monotonic.

Graph 4.22 Age vs Value

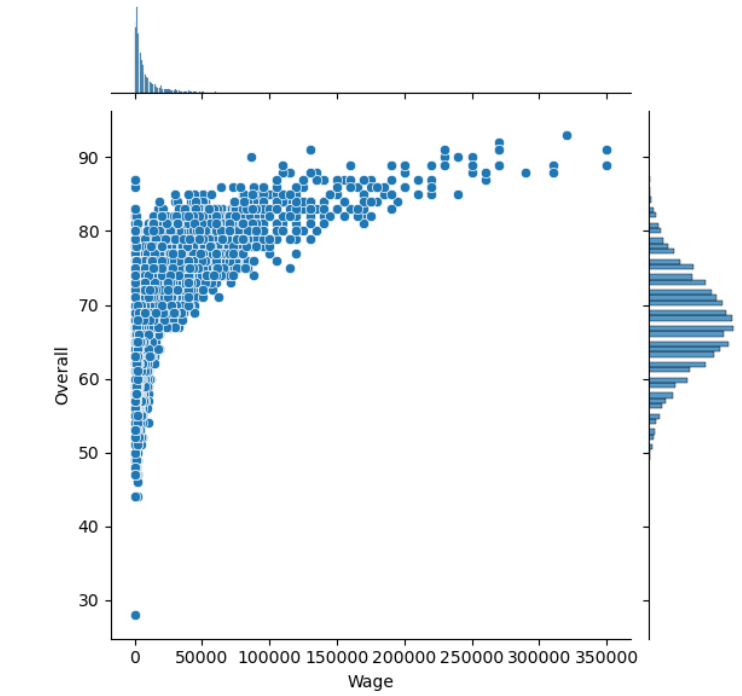
The attributes such as overall rating, age are considered to have considerable impact on wage. Therefore, they have been mapped against wage in the following graphs.

**16. WAGE VS AGE**

Graph 4.23 Wage vs Age

The plot clearly depicts that the players aged 21-29 are highly paid when compared to others. Also, on an average, the wage is 11,338 with a smattering of players getting paid as high as 30 times of the mean value. The distribution therefore is right skewed.

The association of wage with age is not monotonic as there appears to be an optimal range where a player's value peaks. Conversely, the worth of the youngest and oldest players is roughly ten times lower than the maximum.

**17. WAGE VS OVERALL**

The overall rating of a player is proportional to their wages. With increase in their ratings, the wages also tend to increase with a higher concentration of wages lying between 0 and €150000. There are some exceptional players who get paid more than €200000 for higher ratings due to which the wages distribution is highly right skewed here as well.

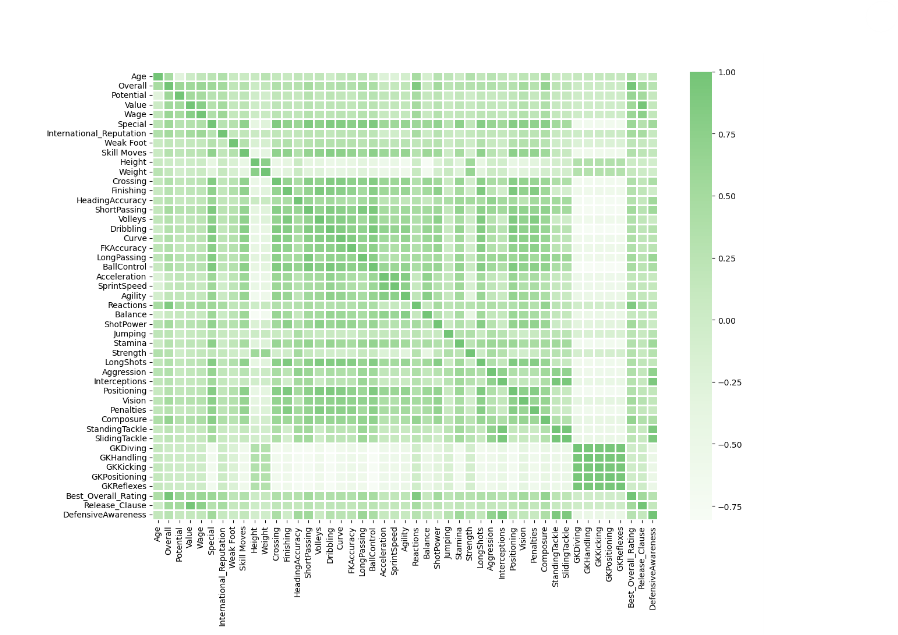
**18. WAGE DISTRIBUTION**

Graphs 3.1 and 3.2 show wage distributions which are highly right skewed. This implies that players who receive low salaries are way more than the number of players who receive high wages. Regardless of their countries,   
position on the field, or physical features, there are football professionals that make salaries that are nearly six times as high as those of contemporary players. This confirms the existence of superstar effect where handful players take home majority of the pay.

Graph 4. Wage vs Overall

**19. HEATMAP**

The correlation between the variables in the DataFrame is displayed in the heatmap below. We can see that there is a strong association between the qualities thought of as skills except for the goalkeepers’ skills which are entirely attributed to the position of goalkeepers and do not apply to the other positions.



Graph 4.25 Heatmap- Attributes correlation

The strength of the correlation between two variables can be interpreted using the shade of the squares where darker the shade, higher the correlation implying that the two variables have a monotonic relationship in that if the value of one variable increases, the value of the other increases as well and vice-versa. For example, apart from being highly correlated themselves in the matrix, columns “Value” and “Release\_Clause” have a very high correlation and therefore, keeping either of them will be enough for the model to work as both of them, if kept together, will lead to misleading results. The shade palette has been given as a reference on the right hand side of the graph.

**4.2.4 DATA CLEANING AND TRANSFORMATION**

**4.2.4.1 Outlier**

Outliers were first treated with capping technique where the 25th and 75th percentiles were used to substitute those values. The error observed after running the baseline model was quite small too. But the issue was the loss of real data which in this case was imperative for our study. The model was run again without treating the outliers and the error was justifiable. Therefore, outliers have not been treated.

**4.2.4.2 Feature selection**

The heatmap was first source of determining the relationship among the attributes at hand and identify the presence of multicollinearity. Multicollinearity is a miracle in statistics when one predictor variable in a multiple retrogression model may be linearly prognosticated with a high degree of delicacy from the others (Brownlee, 2020). Multicollinearity is a concern since it reduces the independent variable's statistical significance (Wu, 2020). The darker the shade, the higher was the level of multicollinearity between the two variables. Based on this these thirteen variables from the dataset were removed. The table has been showed in **APPENDIX A**.

Additionally, to make the work easier the values in the column Best\_Position were regrouped into the four fundamental football positions

Table 4.10 Best\_Position regrouped



**4.2.4.3 One Hot Encoding**

One hot encoding, which improves prognostications and bracket delicacy of a model, is the pivotal process of changing the categorical data variables to be fed to machine and deep literacy algorithms (Brownlee, 2017). Here the categorical variables used were “Preferred foot” and “Best\_Position” for which dummy variables were created.

**4.2.4.4 Feature scaling**

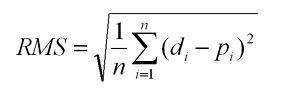
A fashion for normalising the variety of independent variables or features in data is called feature scaling (Roy, 2020). It's generally carried out during the data preprocessing step and is occasionally appertained to as data normalisation in the environment of data processing. It occasionally aids in accelerating algorithmic calculations. Since our dataset had features those varied in scope and units such as weight (in kg), value (in thousands and millions), it was important that machine learning algorithms comprehend these features on the same scale.

The data before and after scaling the datapoints have been mentioned under **APPENDIX B.**

**4.2.5 MODELING USING MACHINE LEARNING ALGORITHMS IN PYTHON**

The data was split into train and test set with the size of the test set of 30% i.e. 5013 test instance and 11697 train instances.

RMSE (root mean squared error) was used as the models; evaluating metrics. The RMSE measures the standard divagation of the crimes that be when a dataset is used to make a vaticination. To put it another way, it's an error in the system used to calculate the delicacy and error rate of a machine learning algorithm used to break a regression problem.. The model fits the data better when the RMSE value is smaller.



Σ - It indicates the "sum".

di- It represents the predicted value for the ith

pi- It represents the predicted value for the ith

n - It represents the sample size.

We have used grid search for most of the models. It is the simplest hyperparameter tuning algorithm. In essence, we created a discrete grid within the range of the hyperparameters, after which, we experimented with all possible grid value combinations and cross-validated some performance indicators. The best setting of values for the hyperparameters was the point on the grid that maximised the average value during cross-validation.

**4.2.5.1 BASELINE MODEL**

The initial straightforward modeling attempt, known as a baseline model, gives us a baseline measure that can be used as a point of reference as we continue to improve the model. This foundational model is a rule-based model, although it could also be a straightforward machine learning model (Allwright, 2022).

As a general guideline, while dealing with regression problems, one should develop baseline models that anticipate the median or mean of the output of the training data.

To create a baseline model, we have used median to be predicted for all test instances found in the training data. The median of wage is 4000. The RMSE of the baseline model is 22668.58 which meant that the wage average error in the dataset is 22668.58 hence, is not a good value.

**4.2.5.2 LINEAR REGRESSION**

A variable's value can be prognosticated using direct regression analysis grounded on the value of another variable. The dependent variable is the one you want to be efficiently forecasted. The independent variable is the one you are using to make a vaticination about the value of the other variable (IBM, 2022).

With the help of one or further independent variables that can most directly prognosticate the value of the dependent variable, this type of analysis calculates the portions of the direct equation. The differencePs between anticipated and factual affair values are minimised by linear regression by fitting a straight line. The best-fit line for a set of paired data can be set up using straightforward direct retrogression calculators that employ the" least places" fashion. The value of X ( the dependent variable) is also estimated using Y (independent variable).

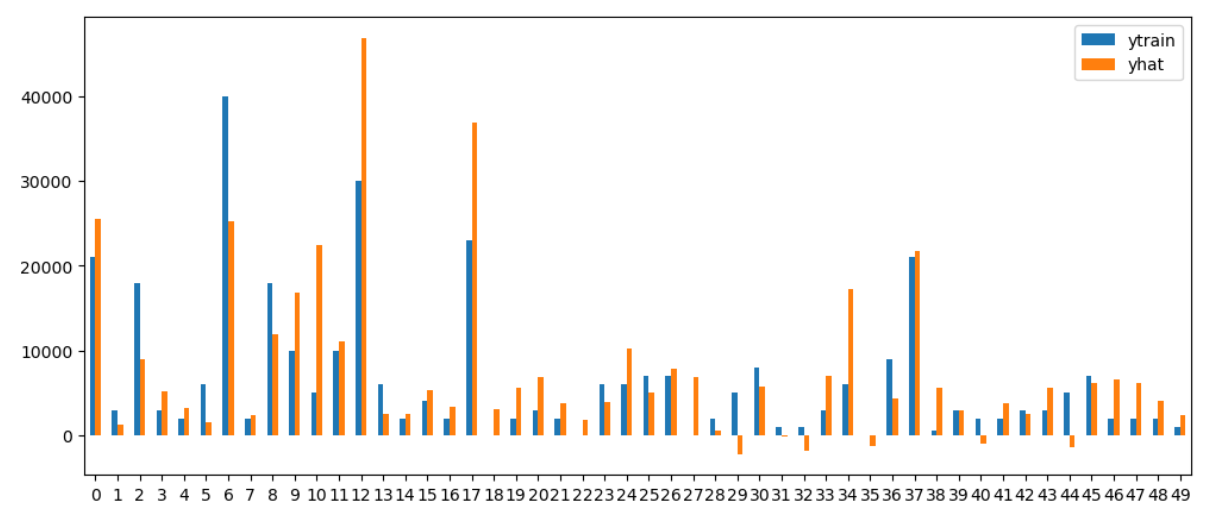
On the train set an *RMSE of 11442.27 was obtained whereas on the test set it was 11793.50*. Taking a look at the r-square, the value of 71.48% was a pretty good result.

An ideal model frequently exhibits identical Train and Test Errors. Therefore, difference percentage was used to find if the model was overfitting or not which, given by:

% difference = 100 \* ((Train score RSME/ Test score RSME)/ Train score RSME)

Here, the difference was 2% only which was very low. Hence the model was not over fitting.

The graph below shows the first 50 instances of the forecasted and actual values. The predictions are shown by ‘yhat’ and the actual values by ‘ytrain’.



Graph 4.26 Actual vs Predicted (first 50 instances)

**4.2.5.3 Decision Tree**

According to Geek for Geeks website, Decision Tree is a method for making decisions that employs a tree structure that resembles a flowchart or is a model of decisions and every conceivable consequence, including outcomes, input costs, and utility.

The supervised learning algorithms group includes the decision-tree algorithm. It works with output variables that can be either categorised or continuous. The nodes have either: the branches/edges represent the node's outcome, and the nodes have

Conditions [Decision Nodes]   
Result [End Nodes]

Based on the validity or falsity of the assertion represented by the branches and edges, a conclusion is made.

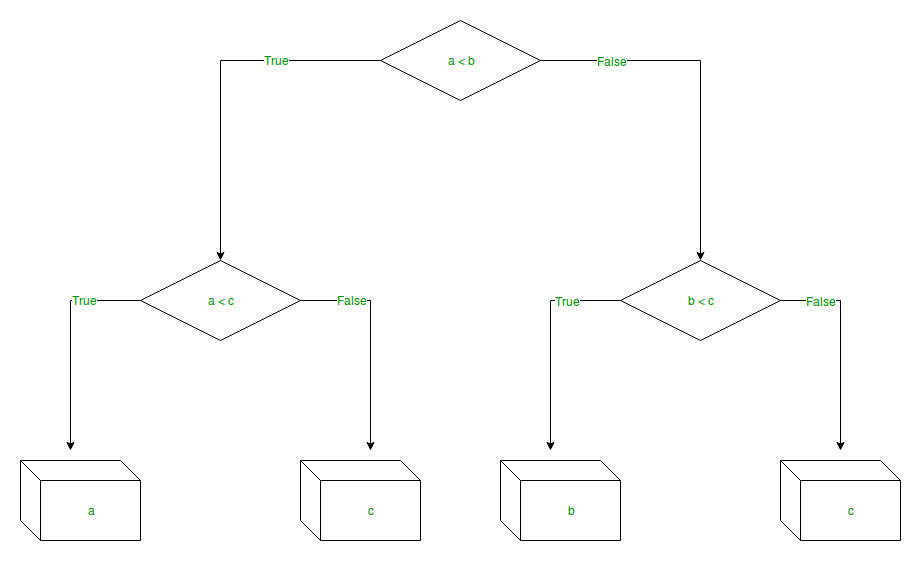


Figure 4.2 Decision Tree

By examining an attribute's characteristics, decision tree regression trains a model in the shape of a tree to forecast data in the future and produce valuable continuous output. The absence of discrete output, or output that is not simply represented by a discrete, well-known set of numbers or values, is referred to as continuous output.

For the decision tree regressor we have used 3 hyperparameters with values:

max\_depth : [1,3,5,7,]

min\_samples\_leaf :[4,5,6,7]

min\_samples\_split: [10, 20, 40]

Therefore, 48 models were created and tested using cross-validation as part of the grid search. The table below shows the combination of hyperparameters along with scores on train and test set. There is another column called the %diff which aids in understanding if the model is overfitting.

Table 4.11 RMSE on Hyperparameter combination- Decision Tree

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_s... | 10269.84869 | 11663.89307 | -13.574147 |
| {'max\_depth': 5, 'min\_samples\_leaf': 4, 'min\_s... | 10265.8477 | 11670.61887 | -13.683928 |
| {'max\_depth': 5, 'min\_samples\_leaf': 4, 'min\_s... | 10015.29612 | 11726.1259 | -17.082169 |
| {'max\_depth': 5, 'min\_samples\_leaf': 6, 'min\_s... | 10271.34501 | 11735.08539 | -14.250718 |
| {'max\_depth': 5, 'min\_samples\_leaf': 7, 'min\_s... | 10267.51931 | 11752.80635 | -14.46588 |
| {'max\_depth': 7, 'min\_samples\_leaf': 7, 'min\_s... | 9892.651374 | 11758.10759 | -18.856989 |
| {'max\_depth': 5, 'min\_samples\_leaf': 6, 'min\_s... | 10105.68494 | 11762.35051 | -16.393402 |
| {'max\_depth': 5, 'min\_samples\_leaf': 6, 'min\_s... | 10394.93975 | 11764.66452 | -13.176842 |
| {'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_s... | 10394.89426 | 11765.44452 | -13.184841 |
| {'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_s... | 10029.92203 | 11778.85128 | -17.437117 |
| {'max\_depth': 5, 'min\_samples\_leaf': 7, 'min\_s... | 10115.17125 | 11779.09295 | -16.449763 |
| {'max\_depth': 5, 'min\_samples\_leaf': 4, 'min\_s... | 10395.75165 | 11780.89648 | -13.324143 |
| {'max\_depth': 7, 'min\_samples\_leaf': 6, 'min\_s... | 9873.499278 | 11781.40806 | -19.323532 |
| {'max\_depth': 5, 'min\_samples\_leaf': 7, 'min\_s... | 10396.85013 | 11789.3682 | -13.393653 |
| {'max\_depth': 7, 'min\_samples\_leaf': 5, 'min\_s... | 9871.975083 | 11798.45275 | -19.514612 |
| {'max\_depth': 7, 'min\_samples\_leaf': 7, 'min\_s... | 9641.540666 | 11873.00578 | -23.144279 |
| {'max\_depth': 7, 'min\_samples\_leaf': 4, 'min\_s... | 9847.403628 | 11905.11047 | -20.895933 |
| {'max\_depth': 7, 'min\_samples\_leaf': 5, 'min\_s... | 9597.292538 | 11922.80851 | -24.230959 |
| {'max\_depth': 7, 'min\_samples\_leaf': 7, 'min\_s... | 9367.566364 | 11955.64293 | -27.628057 |
| {'max\_depth': 7, 'min\_samples\_leaf': 6, 'min\_s... | 9606.270513 | 11965.87003 | -24.563118 |
| {'max\_depth': 7, 'min\_samples\_leaf': 6, 'min\_s... | 9272.740374 | 11984.57087 | -29.245189 |
| {'max\_depth': 7, 'min\_samples\_leaf': 4, 'min\_s... | 9548.790076 | 12041.87419 | -26.108901 |
| {'max\_depth': 3, 'min\_samples\_leaf': 7, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 7, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 7, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 4, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 4, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 5, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 4, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 5, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 6, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 6, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 6, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 3, 'min\_samples\_leaf': 5, 'min\_s... | 11460.19429 | 12070.04687 | -5.321485 |
| {'max\_depth': 7, 'min\_samples\_leaf': 5, 'min\_s... | 9134.323993 | 12170.1213 | -33.235052 |
| {'max\_depth': 7, 'min\_samples\_leaf': 4, 'min\_s... | 9104.804431 | 12329.46224 | -35.417101 |
| {'max\_depth': 1, 'min\_samples\_leaf': 4, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 4, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 5, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 5, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 5, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 7, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 6, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 6, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 7, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 7, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 6, 'min\_s... | 16893.79856 | 17183.30991 | -1.713714 |
| {'max\_depth': 1, 'min\_samples\_leaf': 4, 'min\_s... | 16893.79856 | 17183.30991 | -1.71371 |

On the training set (10269) as opposed to the validation set, the top-performing model displayed a RMSE (11663). This demonstrates that the model overfitted the training set. By placing restrictions on the learned trees, the overfitting may be lessened and the model could be enhanced even further (for example, enforcing a maximum depth of the learned trees). We emphasise that different configurations of the hyperparameters improve performance. As the max depth parameter decreases the risk of overfitting decreases. But with very low values of max depth, error also increases. As seen from the table, max depth of 3 and min sample leaf of 7 gives us a good result with very less evidence of overfitting.

*RSME on the test set was 12046.80*

**4.2.5.4 RANDOM FOREST**

The bootstrapping Random Forest method creates a large number of randomly selected decision trees from the data by combining ensemble learning techniques with the decision tree architecture. A new result is obtained by averaging the results, which typically yields precise forecasts and classifications. (Yiu, 2019).

Ensemble learning is the process of using multiple models that have all been trained on the same data and averaging their results to get a more precise prediction or classification.When bootstrapping, subsets of a dataset are chosen at random over a predetermined number of repetitions and variables (Yiu, 2019). These results are then averaged to produce a more effective result. Bootstrapping is an illustration of an ensemble model in use.

For the random forest regressor, two hyperparameters were used, with values:

n\_estimators: [10, 30,50]

max\_depth: [4, 6, 8, None]

In light of this, 12 models were developed and evaluated by cross-validation as part of the grid search. The table below displays the combination of hyperparameters as well as results from the train and validation sets.

Table 12 RMSE on Hyperparameter combination- Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'max\_depth': 8, 'n\_estimators': 50} | 8011.922352 | 10655.07707 | -32.990269 |
| {'max\_depth': 6, 'n\_estimators': 50} | 9030.863088 | 10677.99222 | -18.238889 |
| {'max\_depth': 8, 'n\_estimators': 30} | 8046.657341 | 10707.41746 | -33.066651 |
| {'max\_depth': 6, 'n\_estimators': 30} | 9043.950965 | 10710.50396 | -18.427267 |
| {'max\_depth': None, 'n\_estimators': 50} | 4095.928557 | 10715.32836 | -161.609259 |
| {'max\_depth': None, 'n\_estimators': 30} | 4194.486037 | 10793.53797 | -157.32683 |
| {'max\_depth': 4, 'n\_estimators': 50} | 10054.81205 | 10824.38976 | -7.653825 |
| {'max\_depth': 4, 'n\_estimators': 30} | 10058.0166 | 10836.89052 | -7.743812 |
| {'max\_depth': 4, 'n\_estimators': 10} | 10136.95061 | 10970.70961 | -8.224949 |
| {'max\_depth': 6, 'n\_estimators': 10} | 9136.969354 | 10981.3571 | -20.18599 |
| {'max\_depth': 8, 'n\_estimators': 10} | 8202.764508 | 10991.37214 | -33.995949 |
| {'max\_depth': None, 'n\_estimators': 10} | 4741.806936 | 11284.99163 | -137.98926 |

The top-performing model had a considerably lower RMSE on the training set (8011) compared to the validation set (10655). The model was shown to have overfit the training set in this case. When the max depth parameter is set to "none" or when it is large (e.g., 8) there is a significant danger of overfitting. There are combinations such as "max depth": 4 and "n estimators": 50 that do not exhibit overfitting.

*The Result of RSME on the test set is 10660.37*.

**4.2.5.5 SUPPORT VECTOR MACHINE**

The objective of the support vector machine algorithm is to locate a hyperplane in an N-dimensional space that categorises the data points precisely. Here, N stands for the number of features

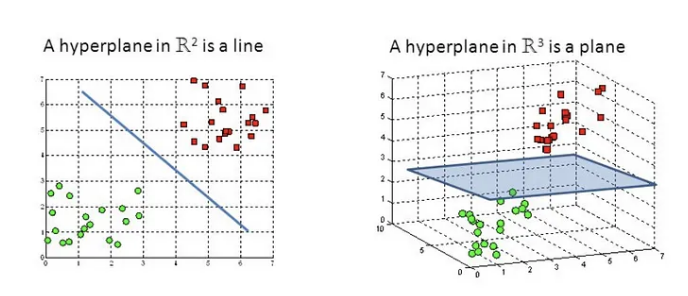


Figure 4.2 Hyperplane in 2D and 3D

The two groups of data points can be divided using a variety of different hyperplanes. Hyperplanes, which serve as decision boundaries, aid in classifying the data points. Our objective is to find a plane with the largest margin, or the largest distance between data points from the two classes. Maximizing the margin distance adds some support and boosts classification confidence for upcoming data points. (Yadav, 2018).

For support vector regressor, two hyperparameters with the following values were used:

C: [1.0,10,100]

gamma': ["scale","auto",0.01,0.1,10]

15 models were created in response to this and tested by cross-validation as part of the grid search. The outcomes from the train and validation sets have been shown in the table below along with the combination of hyperparameters.

Table 4.13 RMSE on Hyperparameter combination- SVM

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'C': 100, 'gamma': 0.01} | 18428.76848 | 18403.67646 | 0.136157 |
| {'C': 100, 'gamma': 'scale'} | 19040.08138 | 19023.44462 | 0.087378 |
| {'C': 100, 'gamma': 'auto'} | 19040.01277 | 19023.51967 | 0.086623 |
| {'C': 10, 'gamma': 0.01} | 21306.49678 | 21271.64327 | 0.163582 |
| {'C': 100, 'gamma': 0.1} | 21347.23151 | 21347.10382 | 0.000598 |
| {'C': 10, 'gamma': 'scale'} | 21493.86489 | 21460.44556 | 0.155483 |
| {'C': 10, 'gamma': 'auto'} | 21494.09013 | 21460.70339 | 0.15533 |
| {'C': 10, 'gamma': 0.1} | 22392.97285 | 22361.86336 | 0.138925 |
| {'C': 1.0, 'gamma': 0.01} | 22457.48771 | 22422.36783 | 0.156384 |
| {'C': 1.0, 'gamma': 'auto'} | 22486.50402 | 22451.48875 | 0.155717 |
| {'C': 1.0, 'gamma': 'scale'} | 22486.51877 | 22451.49876 | 0.155738 |
| {'C': 1.0, 'gamma': 0.1} | 22636.59354 | 22601.46933 | 0.155166 |
| {'C': 100, 'gamma': 10} | 22611.24557 | 22617.96067 | -0.029698 |
| {'C': 10, 'gamma': 10} | 22660.59139 | 22629.21181 | 0.138476 |
| {'C': 1.0, 'gamma': 10} | 22665.54189 | 22630.33993 | 0.15531 |

It can be inferred from the table that none of the models created overfit. The difference percentage is consistent but the RSME on train and validation is way more than that in decision tree, random forest and is close to baseline model. The top-performing model showed considerably the same RMSE on the training set (18428) compared to the validation set (18403). Hence, there is no overfitting but the result of models formed by combination of parameters is very bad.

*The Result of RSME on the test set is 17674.60.*

**4.2.5.6 ADA BOOST**

AdaBoost, also known as Adaptive Boosting, is a machine literacy system used in an ensemble setting. Decision trees with one position, or Decision trees with only one split, are the most popular algorithm used with AdaBoost. This algorithm creates a model while assigning each datapoint an equal weight. also, it gives points that were inaptly categorised larger weights. The coming model now gives further weight to all the points with advancedweights.However, it'll continue to train model until a low error is reported. (Saini, 2021).

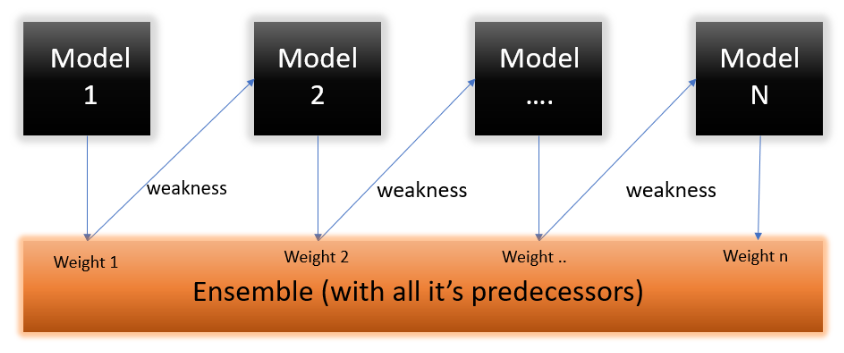


Figure 4.3 Ada Boost model

For ADA boost, three hyperparameters were used with values:

n\_estimators: [3,10,20,50]

learning\_rate: [0.01,0.1,1]

loss: ['linear',’square’,'exponential']

As part of the grid search, 36 models will be developed in response to this and cross-validated. The results from the train and validation sets as well as the combination of hyperparameters are displayed in the table below.

Table 4.14 RMSE on Hyperparameter combination- ADA Boost

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'learning\_rate': 0.1, 'loss': 'exponential', ... | 10513.90374 | 11282.40324 | -7.309364 |
| {'learning\_rate': 0.1, 'loss': 'linear', 'n\_es... | 10523.01968 | 11283.13417 | -7.22335 |
| {'learning\_rate': 0.1, 'loss': 'square', 'n\_es... | 10564.19426 | 11321.08369 | -7.164668 |
| {'learning\_rate': 0.01, 'loss': 'exponential',. | 10801.02038 | 11361.56218 | -5.189712 |
| {'learning\_rate': 0.01, 'loss': 'square', 'n\_e... | 10794.49462 | 11370.34563 | -5.334673 |
| {'learning\_rate': 0.1, 'loss': 'square', 'n\_es... | 10709.34821 | 11396.78098 | -6.418997 |
| {'learning\_rate': 0.01, 'loss': 'linear', 'n\_e... | 10788.60831 | 11402.92345 | -5.694109 |
| {'learning\_rate': 0.1, 'loss': 'linear', 'n\_es... | 10707.87275 | 11430.74003 | -6.750802 |
| {'learning\_rate': 1, 'loss': 'linear', 'n\_esti... | 10671.07711 | 11436.69929 | -7.174741 |
| {'learning\_rate': 0.01, 'loss': 'exponential',... | 10952.29893 | 11439.41576 | -4.447622 |
| {'learning\_rate': 0.1, 'loss': 'exponential', ... | 10689.66964 | 11441.74589 | -7.035543 |
| {'learning\_rate': 0.1, 'loss': 'linear', 'n\_es... | 10857.91085 | 11465.06263 | -5.591792 |
| {'learning\_rate': 0.1, 'loss': 'exponential', ... | 10884.71404 | 11495.40885 | -5.610573 |
| {'learning\_rate': 1, 'loss': 'exponential', 'n... | 10584.65055 | 11501.5116 | -8.662176 |
| {'learning\_rate': 0.01, 'loss': 'square', 'n\_e... | 10930.36724 | 11513.56015 | -5.335529 |
| {'learning\_rate': 0.01, 'loss': 'linear', 'n\_e... | 10940.4605 | 11521.57272 | -5.311588 |
| {'learning\_rate': 0.1, 'loss': 'square', 'n\_es... | 10883.70895 | 11538.58656 | -6.017044 |
| {'learning\_rate': 0.01, 'loss': 'exponential',... | 11032.16288 | 11586.2496 | -5.022467 |
| {'learning\_rate': 1, 'loss': 'square', 'n\_esti... | 10627.80195 | 11587.16286 | -9.026899 |
| {'learning\_rate': 0.01, 'loss': 'linear', 'n\_e... | 11020.72526 | 11606.22197 | -5.312688 |
| {'learning\_rate': 0.01, 'loss': 'square', 'n\_e... | 11061.6694 | 11682.66907 | -5.613978 |
| {'learning\_rate': 1, 'loss': 'exponential', 'n... | 10963.20186 | 11748.02832 | -7.158734 |
| {'learning\_rate': 1, 'loss': 'square', 'n\_esti... | 11177.63981 | 11822.09414 | -5.765567 |
| {'learning\_rate': 0.1, 'loss': 'square', 'n\_es... | 11220.17591 | 11874.56105 | -5.832218 |
| {'learning\_rate': 0.1, 'loss': 'linear', 'n\_es... | 11382.78679 | 11900.96603 | -4.552306 |
| {'learning\_rate': 1, 'loss': 'linear', 'n\_esti... | 11087.17024 | 11958.72505 | -7.860931 |
| {'learning\_rate': 1, 'loss': 'square', 'n\_esti... | 10903.36489 | 11981.7192 | -9.890106 |
| {'learning\_rate': 0.01, 'loss': 'linear', 'n\_e... | 11390.09297 | 11992.9777 | -5.293062 |
| {'learning\_rate': 1, 'loss': 'linear', 'n\_esti... | 11136.29109 | 11993.22987 | -7.695011 |
| {'learning\_rate': 0.01, 'loss': 'exponential',... | 11370.01786 | 12026.84166 | -5.776805 |
| {'learning\_rate': 0.01, 'loss': 'square', 'n\_e... | 11371.07214 | 12048.76138 | -5.959766 |
| {'learning\_rate': 1, 'loss': 'exponential', 'n... | 11169.86946 | 12146.8571 | -8.746634 |
| {'learning\_rate': 0.1, 'loss': 'exponential', ... | 11468.31662 | 12209.21779 | -6.460418 |
| {'learning\_rate': 1, 'loss': 'linear', 'n\_esti... | 19566.31329 | 20093.58729 | -2.694805 |
| {'learning\_rate': 1, 'loss': 'exponential', 'n... | 19760.82184 | 20350.08356 | -2.98197 |
| {'learning\_rate': 1, 'loss': 'square', 'n\_esti... | 20818.52538 | 21456.7795 | -3.06579 |

When compared to the validation set (11282), the top-performing model had a significantly lower RMSE on the training set (10513). In a small number of cases, it was discovered that the model had somewhat overfit the training set. As a result, the model is performing well on the training and validation sets, and the results on both sets are also extremely near, reducing the overfit. The model's parameters are "learning rate": 0.01 and "loss": "exponential" or "square." There are models with various combinations of hyperparameters in which overfitting is essentially non-existent, yet the RSME on the training and validation set is extremely high, indicating that the model is not functioning properly.

*The Result of RSME on the test set is 11064.80.*

**4.2.5.7 XGBOOST**

The gradient boosted trees approach is widely used and well implemented in open-source software called XGBoost. Gradient boosting is a supervised learning process that combines the predictions of weaker, simpler models to attempt to properly predict a target variable (Amazon, 2022).

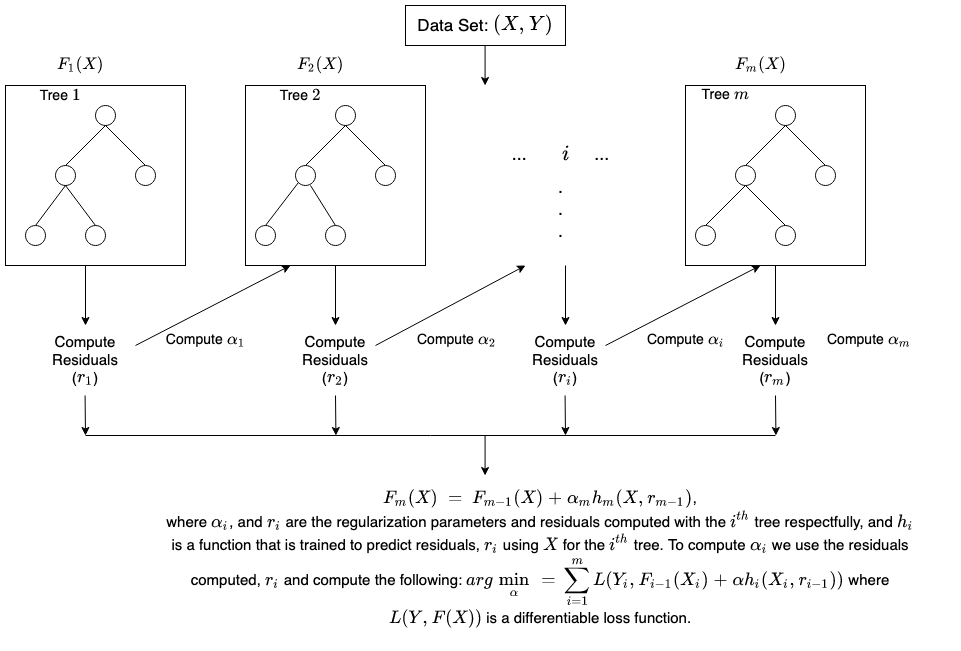


Figure 4.4 XGBoost Model

Regression trees serve as the weak learners when utilising gradient boosting for regression, and each one of them associates each input data point with a leaf that holds a continuous score. With a convex loss function( grounded on the difference between the prognosticated and target labors) and a penalty term for model complexity, XGBoost minimises a regularised( L1 and L2) objective function. Adding new trees that forecast the residuals or errors of earlier trees, which are then integrated with earlier trees to produce the final prediction, is how the training process is carried out iteratively.

For XG boost regressor, three hyperparameters with values:

n\_estimators: [100, 500]

max\_depth: [3,5]

min\_child\_weight: [5,8,12]

12 models were generated in accordance to this and cross-validated as part of the grid search. The results from both the train and validation sets in addition to the pairing of hyperparameters are shown in the table below.

Table 4.15 RMSE on Hyperparameter combination- XG Boost

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'max\_depth': 3, 'min\_child\_weight': 8, 'n\_est... | 7129.855634 | 10709.10569 | -50.200877 |
| {'max\_depth': 3, 'min\_child\_weight': 12, 'n\_es... | 7267.917576 | 10814.7523 | -48.801251 |
| {'max\_depth': 3, 'min\_child\_weight': 5, 'n\_est... | 7043.734977 | 10897.2264 | -54.70807 |
| {'max\_depth': 3, 'min\_child\_weight': 8, 'n\_est... | 4178.070796 | 11010.08152 | -163.520703 |
| {'max\_depth': 5, 'min\_child\_weight': 8, 'n\_est... | 4569.587786 | 11090.25956 | -142.697155 |
| {'max\_depth': 5, 'min\_child\_weight': 12, 'n\_es... | 4864.182995 | 11092.70108 | -128.048597 |
| {'max\_depth': 5, 'min\_child\_weight': 5, 'n\_est... | 4311.022508 | 11186.24593 | -159.48011 |
| {'max\_depth': 3, 'min\_child\_weight': 12, 'n\_es... | 4279.941389 | 11257.43787 | -163.027851 |
| {'max\_depth': 5, 'min\_child\_weight': 8, 'n\_est... | 1091.4656 | 11300.60657 | -935.360764 |
| {'max\_depth': 3, 'min\_child\_weight': 5, 'n\_est... | 4099.54707 | 11313.28351 | -175.96423 |
| {'max\_depth': 5, 'min\_child\_weight': 12, 'n\_es... | 1246.197151 | 11356.97237 | -811.330311 |
| {'max\_depth': 5, 'min\_child\_weight': 5, 'n\_est... | 984.621944 | 11387.58505 | -1056.5439 |

On the training set (7129) as opposed to the validation set, the top-performing model displayed a significantly reduced RMSE (10709). This demonstrates that the model overfit the training set. Irrespective of different combinations of the max depth and min child weight parameters all the models are overfitting. Hence XG boost is badly performing on the train and validation data set.

*The Result of RSME on the test set is 11113.23*.

**4.2.5.8 K-NEAREST NEIGHBOUR**

For KNN, two hyperparameters were used with values:

n\_neighbors: [2,3,4,5,6]

weights: ['uniform','distance']

In accordance with this, 10 models were created and cross-validated as part of the grid search. The table below displays the results from the pairing of hyperparameters as well as the train and validation sets (Amazon, 2022).

Table 4.16 RMSE on Hyperparameter combination- KNN

|  |  |  |  |
| --- | --- | --- | --- |
| **Params** | **mean\_train\_score** | **mean\_test\_score** | **diff, %** |
| {'n\_neighbors': 6, 'weights': 'distance'} | 0.001731 | 12090.82082 | -698508800.00 |
| {'n\_neighbors': 6, 'weights': 'uniform'} | 10114.64285 | 12123.13532 | -19.86 |
| {'n\_neighbors': 5, 'weights': 'distance'} | 0.001422 | 12201.22901 | -857896900.00 |
| {'n\_neighbors': 5, 'weights': 'uniform'} | 9852.88382 | 12217.93649 | -24.00 |
| {'n\_neighbors': 4, 'weights': 'distance'} | 0.00113 | 12425.70835 | -1099798000.00 |
| {'n\_neighbors': 4, 'weights': 'uniform'} | 9503.113554 | 12431.74566 | -30.82 |
| {'n\_neighbors': 3, 'weights': 'uniform'} | 8981.739088 | 12719.61821 | -41.62 |
| {'n\_neighbors': 3, 'weights': 'distance'} | 0.000834 | 12733.56375 | -1526338000.00 |
| {'n\_neighbors': 2, 'weights': 'uniform'} | 7888.713883 | 13512.22174 | -71.29 |
| {'n\_neighbors': 2, 'weights': 'distance'} | 0.000541 | 13522.14582 | -2.50 |

When compared to the validation set (10114), the top-performing model had a very good RMSE on the training set (12090). In almost every case, it was discovered that the model had overfit the training set. As a result, the model is performing worst in most of the cases, on the validation set and the results on validation sets are extremely high, increasing the overfit. In few cases where the RSME on the training set is 0 suggesting that the average wage predicted was ideal but, on the validation set error is way to large. Hence, this model is not fit to use.

*The Result of RSME on the test set is 12436.93*.

**4.3 CLUB REVENUE ANALYSIS- GRAPHICAL ANALYSIS USING TABLEAU**

Furthermore, an additional analysis was conducted on how club revenues affect the player wages. The leagues analysed here were-  
1. English Premiere League (England)  
2. Bundesliga (Germany)  
3. La Liga (Spain)  
4. Ligue 1 (France)  
5. Series A (Italy)  
6. English Championship (Stoke City) [For Part B only]

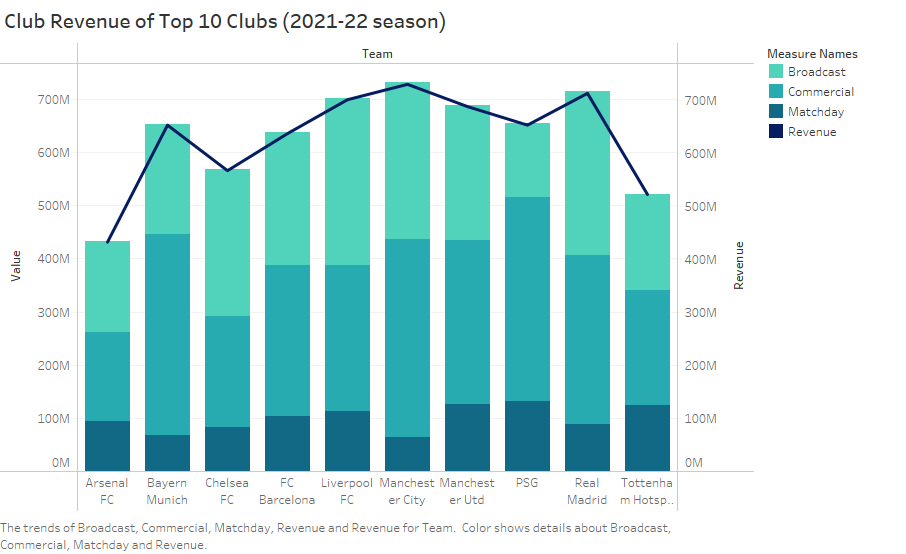
Using tableau, the target variable - **wage was analyzed in two steps**. First, the trend among the top 10 clubs across the five top leagues was studied to study the correlation between the revenue earned by the clubs and the percentage of the revenue that they can afford to spend as players’ salaries. Next, two clubs were taken from England’s two different divisions of football league in order to understand how two leagues within a nation of one of the top football talents function in terms of funds. The table of raw data has been captured under **APPENDIX C**.

**4.3.1 ANALYSIS OF THE TOP 10 CLUBS ACROSS FIVE MAJOR LEAGUES**As per Deloitte’s Money league study report for 2021-22 season, 20 top clubs were identified. But here 10 of those clubs were selected who were at the top of the Champions league that season. Manchester City topped the list with the highest revenue. It is interesting to note that it individually had the highest revenue grossed from all the three sources as well.

**4.3.1.1 CLUB REVENUE OF TOP 10 CLUBS**

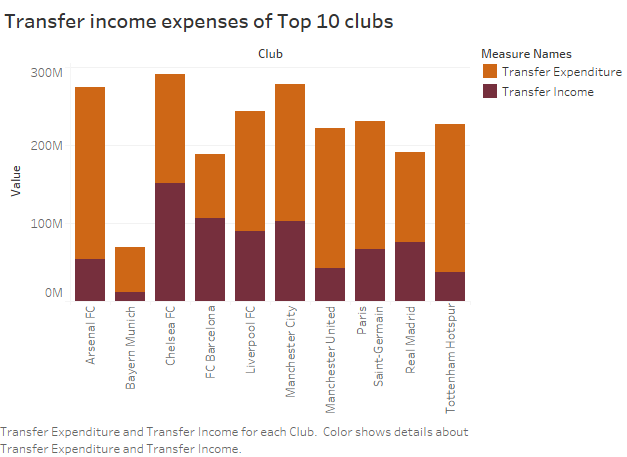
The graph explains the humungous revenues that these clubs have been generating through sponsors, broadcasting rights and matchday ticket and merchandise sales due to heavy fan following. The stacked bars represent broadcast (top), commercial (middle) and matchday (bottom) revenues vertically.

The line graph represents the revenue of each team which is the sum of these three sources. It is very clear that the top tier clubs do not have any concern in terms of availing funds and therefore, they end up spending way too much on their players which can be seen in the graph.



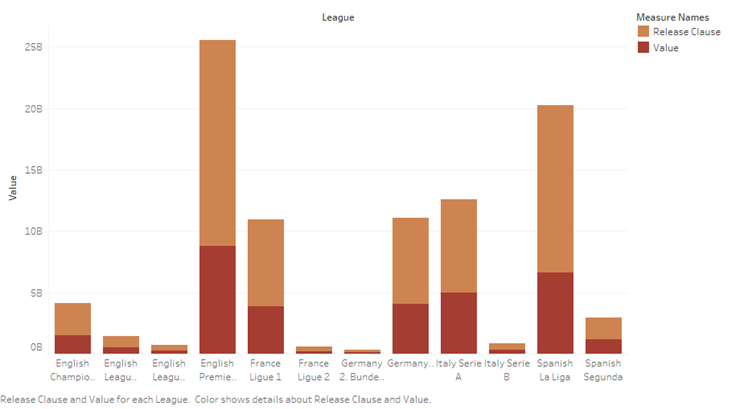
Graph 4.27 Club Revenue of Top 10 clubs

**4.3.1.2 TRANSFER INCOME VS TRANSFER EXPENDITURE**



Graph 4.28 Transfer expenses of top 10 clubs

With the allocated transfer income, most of these clubs have maintained a positive balance sheet but teams such as Liverpool, Paris Saint Germain and Bayern Munich have spent a lot on transfers which imply that these clubs with high revenues and sponsors go out of the way to acquire players from other teams even before their contracts expire. Bayern Munich being the only club from Bundesliga and a representation of the league explains how the salary caps in Germany are not really working for the leagues due to which clubs still end up paying way more than required (Frick, 2008).  
  
**4.3.1.3 RELEASE CLAUSE VS VALUE**

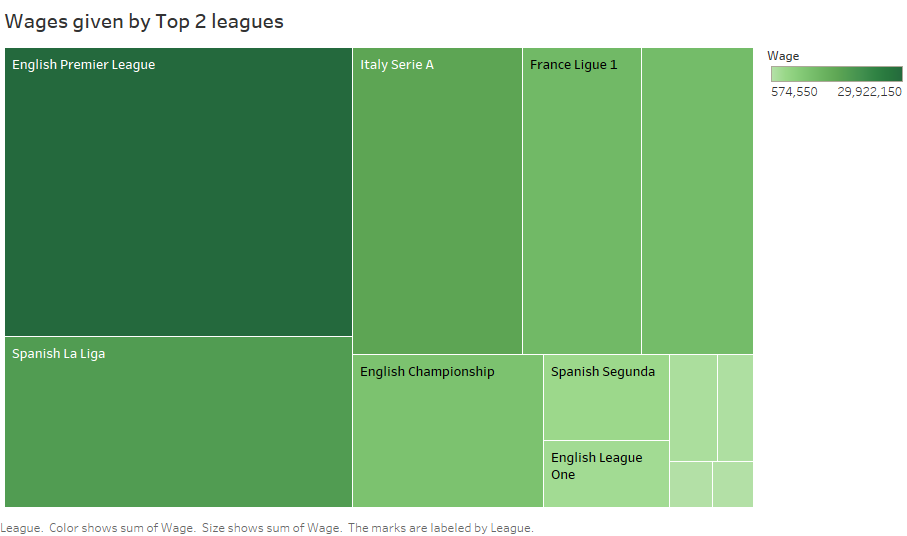


Graph 4.29 Release Clause vs Value

As proven from our machine learning analysis, value and release clause are directly proportional to each other. Here the graph depicts the sum of release clauses of the players in every league stacked on the sum of the values of the players in that league. For the comparison, two top division leagues were selected from the five countries. As expected, the first division teams of all the countries possess a big chunk of the market when it comes to owning the highly valued players. English Premiere League leads here as well, followed by Spanish La Liga and Italian Serie A.

**4.3.1.4 HEATMAP- WAGES PAID TO THE PLAYERS**

Finally, the map below shows the comparison of the amount of salaries that the top two leagues from the countries England, France, Germany, Italy and Spain pay their players in total. English Premiere League being one of the most elite leagues of the world leaves no stone unturned when paying their salaries. Germany, due to some caps applied by the government certainly falls a little behind when compared to the others but it still has its best players such as Robert Lewandowski earning as much as Kevin de Bruyne and Christiano Ronaldo.



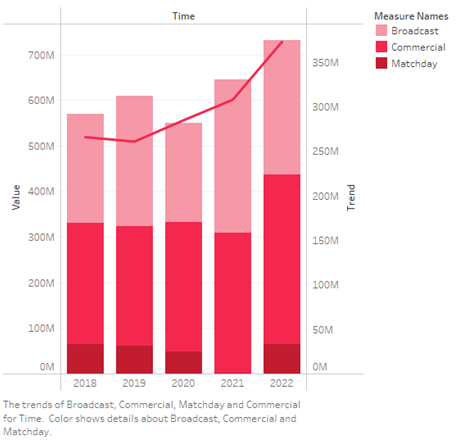
Graph 4.30 HEATMAP- Highest Wage paying Clubs

**4.3.2 MANCHESTER UNITED VS STOKE CITY- ANALYSIS OF TEAMS OF DIFFERENT ENGLISH LEAGUES**

The snapshots of datasets for SCFC and MUFC have been captured under **APPENDIX D** and **APPENDIX E** respectively.

**4.3.2.1 Manchester United (MUFC)**

Initially founded in 1878, the club is currently owned by Malcolm Glazer’s six children through an investment company Red Football Ltd. (Football History). The club has seen managers like Sir Alex Ferguson and built the careers of players such as Rooney, Eric Cantona, David Beckham and Christiano Ronaldo. According to the club’s website, over the time it has garnered 1.1 billion fans worldwide supporting the clubs throughout. It has sponsors with huge pockets such as Adidas, Team Viewer, Apollo Tyres, etc. The club plays in England’s first division, English Premiere League and is frequently one of the league toppers to qualify for Champions league. It has won the title thrice with the last win registered in 2008.



Graph 4.31 Club revenue of Manchester United

**4.3.2.2 Stoke City (SCFC)**

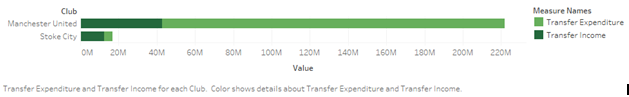
Founded as Stoke Ramblers by Henry Almond in the 1863, it is currently owned by Bet365 Group Ltd. Stoke City currently plays in the English Championship (Second Division league). It last played in the EPL in the 2017-18 season and the highest spot that it had garnered was 9 till now. Peter Crouch and Stanley Matthews are some of its greatest players of all time. Some of its sponsors are Foodhub, Staffordshire University and Franklyn Financial Management.

Table 4.17 Manchester Utd vs Stoke City revenue and expense data



**4.3.2.3 KEY ANALYSIS**

* Stoke City Football Club's revenue declined in 2021 by £10 million due to COVID, which resulted in match day revenue falling by £4.7 million (97%) to just £121 thousand and commercial revenue falling by £1.9 million (14%) to £12.0 million.
* Broadcasting's decline of £2.9 million (9%) to £28.3 million was partially offset by deferred 2019–20 revenue.
* Although Stoke City only reduced the wage bill by £5 million (8%) from £55 million to £50 million, player amortisation and player impairment saw significant drops of £19 million (64%) and £39 million (91%) from £43 million to £4 million, respectively.
* Other costs increased by £7m (62%) to £19m.
* Stoke City’s last disclosed revenue from sponsorship was around €8.8 million in 2020 which was only 1/36th of what Manchester United had around that time. Manchester United had a whopping €322 million during Covid times
* The total salaries that the clubs had allocated as players’ wages in 2021-22 were €17.11 million (SCFC) and €271.9 million (MUFC)
* When it comes to inflated salaries, SUFC paid gross inflated salary of €27.134 million while MUFC paid €268.6 million.



Graph 4.32 MUFC vs SCFC revenue comparison

The graph above shows the huge difference between the transfer income (dark green) and transfer expenditure (light green) of two clubs.

CHAPTER V  
RESULTS AND DISCUSSION

**5.1 PART A- WAGE PREDICTION MODEL**

**5.1.1 RMSE COMPARISON**

The best RMSE results achieved on the training set and on the test set from all the seven models were tabulated for comparison to identify the best model. Given below is a table of best RSME achieved on the training set and on the test set:

Table 5.1 RMSE Comparison



We can infer from the table that ADA boost is one of the top performing models as it showed very low or almost no evidence of overfitting as well as performed well on both the datasets. To support the result further, scatterplots of the actual wages and predicted outputs for all the seven models are shown below-

Table 5.2 SCATTERPLOTS (PREDICTED VS ACTUAL)

|  |  |
| --- | --- |
| LINEAR REGRESSION | DECISION TREE |
| RANDOM FOREST | SUPPORT VECTOR MACHINE |
| XGBOOST | K-NEAREST NEIGHBOUR |

The graph for linear regression shows how well the regression line fits the data points. It can be seen that most of the values were correctly predicted and hence they follow the red line. The best fit line covers almost all of the data points thus implying that the prediction has been successful.

**ADA BOOST**

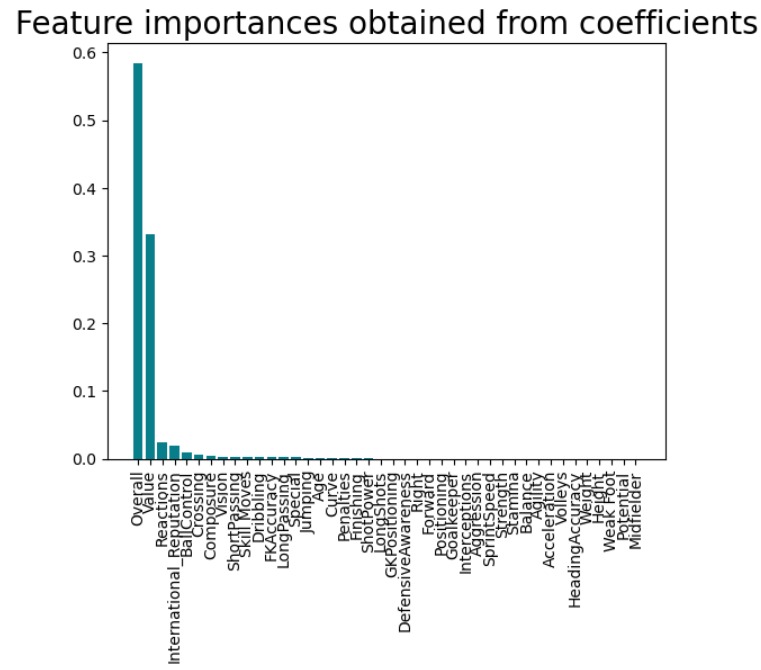
The scatterplot for ADA Boost fits the train set without any overfitting like in the case of K-Nearest Neighbour. The best fit covers the points uniformly.Therefore, ADA Boost is the best model.



Graph 5.1 Actual vs Predicted (ADA Boost)

**5.1.2 FEATURE IMPORTANCE (DATASET LEVEL EXPLANATION)**

Since ADA Boost model was proved to be the best model, a feature importance graph was further plotted to identify the parameters that contributed the most to the target variable in a decreasing order. The feature importance of associated variables is frequently evaluated from the coefficients of linear models. In general, feature usefulness in forecasting a target variable is referred to as feature importance.



Graph 5.2 Feature importance

Based on the results, the significant variables are overall followed by value. Following five attributes are significant enough to be pointed out-

1. **Overall –** The player's in-game stats make up the OVR (Overall Value Rating). Additionally, the position is quite important. OVR and performance are statistically significantly correlated. Better performance results from higher OVR. With a 95% certainty, the performance improvement is between 0.005 and 0.013 extra goals per game every additional OVR-rating point (Sohns, 2021).
2. **Value –** This is the estimated price (€) at which a team can transfer a player's contract to another team is known as the player's market value.
3. **Reactions –** The speed at which a player reacts to things happening around them.

Graph 5.3 MUFC- Actual wage vs Predicted wage against Reactions rating

Graph 5.4 SCFC- Actual wage vs Predicted wage against Reactions rating

The variations in MUFC’s players’ reaction towards the ball and opponents were captured over a range of 59-94 while that of SCFC’s was 40-75 with many players having the same reaction value. Since there was not much variation for the latter, the predicted wages were capped around three major trigger points as evident from the Graph.

1. **International Reputation** - International reputation is a rating system from 1 to 5 which helps in determining the final overall score
2. **Ball Control-** Player's ability to control the ball while in their possession.

The importance of attributes those follow are not significant enough to be compared to the first five attributes.

**5.1.3 ACTUAL VS PREDICTED WAGES**

The predicted and actual weekly wages (in Euros) paid by two English clubs from different leagues were extracted, tabulated and graphically represented to compare the differences between the values.

Graph 5.5 MUFC- Actual Wage vs Predicted wage comparison

This side by side graph uses dark blue bar to represent the actual wage values for Manchester United (English Premiere League) and light blue bar to represent the predicted value for 41 players from the dataset. The model has predicted the wages based on the performance of the players. With increase in overall rating, the wages of the players tend to increase as well.

Graph 5.6 SCFC- Actual Wage vs Predicted wage comparison

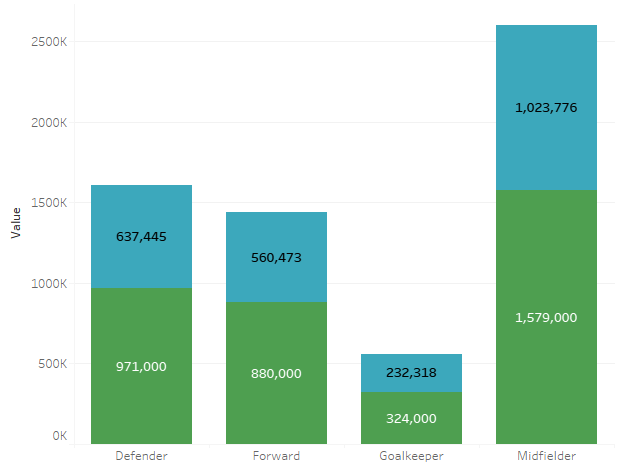
This side by side graph uses blue line to represent the actual wage values for Stoke City (English Championship) and red line to represent the predicted value for 34 players from the dataset. Since majority of the players lie in the low rating range of 50-68, the predicted wages do not reflect a lot a variation in wages with increase in performance. The predicted weekly wage here is €3447 for actual wages €2000 to €12000 for their corresponding 50-68 ratings’ values. It increases again but remains almost constant at around €12,000 for ratings 69-73 before finally spiking up to 74. But justifying such a leap from 12000 to 25000 can only be possible using the second feature called “Value” which happens to be

From both the graphs, it can be concluded that there is a difference between the wages that the players are currently receiving as salary and what they are being paid. The trend in the salaries among different players remains the same though.

|  |  |  |
| --- | --- | --- |
| **CLUB (weekly wage)** | **MUFC** | **SCFC** |
| Actual Wages | 3754000 | 445000 |
| Predicted Wages | 2454012 | 288938.6 |

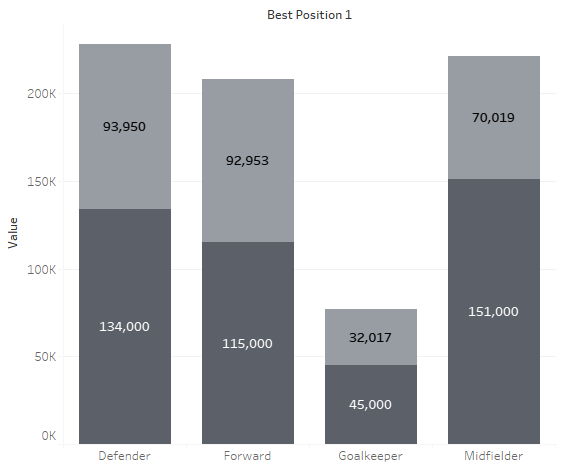
The sum of the market values of the 41 players in MUFC is 1,027,631,000 whereas for the players in SCFC is 60,320,000 Euros.

The wages have been further segregated on the basis of the “Best\_Position” in the following stacked bar chart where the blue stack represents the sum of the predicted wages and green stack represents the sum of the actual wages for MUFC.



Graph 5.7 MUFC- Wage distribution based on Position

Similarly, the wages for SCFC have been divided on the basis of the “Best\_Position” in the following stacked bar chart where the light gray stack represents the sum of the predicted wages and dark gray stack represents the sum of the actual wages.



Graph 5.8 SCFC- Wage distribution based on Position

**5.1.4 INSTANCE LEVEL EXPLANATIONS**

Considering some use cases from the result based on the feature importance obtained above to further elaborate on the findings.

**CASE 1- MUFC- Christiano Ronaldo vs Fred (Overall Rating)**

Both of them play as forwards for the team but Ronaldo with overall rating of 91 gets paid more than twice of Fred whose rating is 81 despite Ronaldo (36) being 8 years older than Fred (28). This shows that **overall rating** is an integral factor in determining a player’s wage.

**CASE 2- MUFC- Bruno Fernandes vs Jaden Sancho**

Finest of the midfielders, Fernandes (26) and Sancho (21) have overall ratings of 88 and 87 respectively but they get paid 250k and 150k weekly. This conforms to the general trend where players between the age of 24 and 33 get paid the highest amount of salaries for both their skills and expertise. Therefore, even though Sancho is young, his goal conversion and assist rate are lower than that of Fernandes.

**CASE 3- PAUL POGBA (International Reputation)**

A 28 year old midfielder, Paul Pogba earns 220K weekly when the predicted salary was just 146.37K leading to a difference of staggering 73.623K euros. A star player, Pogba gets paid way more than his fellow players and a suitable explanation for this is his market value (second-best feature) and international reputation of 4 stars which is way more than them.

Graph 5.9 International Reputation MUFC vs SCFC

Also, international reputation of players

**CASE 4- MUFC H. Mejbri VS S. Shoretire**

Young midfielders aged 18 and 17 respectively, despite with overall ratings of 62 and player positions CAM, the difference in their weekly salaries is noticeable at 6000 and 2000 respectively when the predicted salary based on their ages and ratings is 3448. This shows that the wages process is not very structured in the leagues.

**CASE 5- MUFC- Same rating- 78**

Table 5.3 Case 5- MUFC players comparison with same OVR

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Age** | **Overall** | **Value** | **Best Position** | **Regrouped** | **Wage** | **Predicted Wage** |
| M. Greenwood | 19 | 78 | 30500000 | CAM | Midfielder | 56000 | 49943.3662 |
| Â R. Giggs | 39 | 78 | 1000 | CAM | Midfielder | 60000 | 26954.3438 |
| Juan Mata | 33 | 78 | 9000000 | CAM | Midfielder | 100000 | 31576.6332 |

Amongst the three players, for the same position and overall rating, the predicted values of wages are different. Therefore, based on the second best feature after “overall”, Value is highest for Greenwood and hence needs to be compensated more. Age also is a factor here where Giggs and Mata being 30+ have their value reduced considerably. Therefore, the predictions for MUFC are both performance and popularity based.

**CASE 6- SCFC- Same Rating- 71**

There are four players with overall rating of 71 and all of them play in different position but have the same salary.

Table 5.4 CASE 6- SCFC- Same Rating- 71

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Age** | **Overall** | **Value** | **Best Position** | **Regrouped** | **Wage** | **Predicted wage** |
| H. Souttar | 22 | 71 | 3900000 | CB | Defender | 14000 | 12295.9879 |
| A. Davies | 28 | 71 | 1600000 | GK | Goalkeeper | 14000 | 12295.987 |
| 21Â M. Bauer | 28 | 71 | 1700000 | RB | Defender | 18000 | 12295.9879 |
| T. Smith | 29 | 71 | 1600000 | RWB | Defender | 18000 | 12295.9879 |

Based on the table, it can be inferred that the wage being paid to the players are quite closer to the predicted wage. Also, the predicted wage depends entirely on the players overall rating irrespective of the fact that they play from different “positions” and have different “Value”.

Although, the value for H, Souttar is more than double of others, yet his predicted wage is same as the others because the “value” of Souttar might be higher than his teammates but overall when compared to the other players in the dataset it does not make much of a difference in the predicted values.

The actual wages on the other hand makes sense too in that the age of players also play an important role in their current wages. In the real world, both old and young players are not valued high.

**CASE 7- SCFC vs MUFC (Rating- 74)**

Players with rating 74 in both the clubs are predicted to be paid the same amount. Since the ratings and ages are comparable, they indeed are paid the same which shows that the model predicts uniformly.

Table 5.5 Jones vs Owen



But Owen being the only highest rated player in SCFC, when compared with the highest rated player from MUFC, the results were quite surprising but expected.

Table 5.6 Owen vs Ronaldo



Ronaldo, being the best player in the united’s squad and across the world with the highest OVR, also excels in every way when compared to Owen. This shows the gap between the talent acquisition between the two clubs where the best player of SCFC is only as good as an average player of MUFC

To summarize, if MUFC and SCFC are considered representatives of their respective leagues, the instance level explanations highlight the difference between the wage structures of the clubs. As per feature importance, wage depends majorly on “Overall rating” of the players followed by their “Value”. From the sample space of these two clubs, SCFC has 34 players with overall rating ranging between 50 and 74 whereas MUFC has 41 players with ratings between 60 and 91. While SCFC has a rating as low as 50, its highest rated player, A.M. Owen has a rating of 74 only with highest predicted weekly wage of 24929. MUFC has the lowest rated player, D. Hoogewerf being paid 3447 while the highest rated, Christiano Ronaldo with predicted salary of 236666.667. Therefore, small clubs also have a dearth of talented players.

Table 5.7 MUFC Players Sample

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Age** | **Overall** | **Value** | **Best Position** | **Regrouped** | **Wage** | **Predicted Wage** |
| T. Mengi | 19 | 64 | 1400000 | CB | Defender | 10000 | 3447.924 |
| A. Elanga | 19 | 65 | 1700000 | CAM | Midfielder | 12000 | 3447.9244 |
| L. Grant | 38 | 65 | 80000 | GK | Goalkeeper | 11000 | 3447.924 |
| A. Diallo | 18 | 68 | 3100000 | CAM | Midfielder | 12000 | 3447.9244 |
| P. Jones | 29 | 74 | 3500000 | CB | Defender | 73000 | 24929.5911 |
| Â R. Giggs | 39 | 78 | 1000 | CAM | Midfielder | 60000 | 26954.3438 |
| Juan Mata | 33 | 78 | 9000000 | CAM | Midfielder | 100000 | 31576.6332 |

**5.1.5 SUPERSTAR EFFECT**

Frank and Cook (1995) explored the idea of winner-take-all compensation systems in the economy. "Winner-take-all" is defined as the superstar phenomenon, in which slight variations in performance result in significant variations in rewards. Differences in talent and an inherent property right that the victor can assert are two requirements for superstar salaries. Although it is often more acceptable and understandable for competitors to have different levels of talent, yet the idea that endogenous property rights contribute to superstar wages is more contentious.

While the salaries of the SCFC players are close to the predicted salaries, it is also in alignment with the overall rating of the players where the predicted wages are quite proportional to the overall ratings. But due to less variation in the skill values, the predicted wages have been clustered for many players. Therefore, even if it increases with the predictors’ values, yet the difference in wages cannot be seen until a major performance change occurs. Also, from the Table 5.7, it can be seen that the predicted wages and actual wages of the players in MUFC have a lot of variations. Therefore, there is a rise in wages with change in performance values. Additionally, some players are paid more than needed. With Ronaldo transcending some of the established facts such as regarding age where being 36 year old he is being paid more, it establishes the existence of “superstar effect” which is very common in the big clubs.

The findings align with that of Adler’s study (as per Case 7) where difference in popularity such as that of Ronaldo, is what getting him paid more than the standard. Therefore, it is the popularity of the player leading to superstar effect and not small disparities in talent.

**5.1.6 TOURNAMENT EFFECT**

**Although not very prominent, but still tournament effect exists based on the** cases 2,3 and 7 where players with high international reputation such as Pogba gets paid much more than other players just because he has a 4 star rating on reputation. He did not perform well last season with hardly one goal and therefore has been loaned out this season. Also, Bruno Fernandes gets paid more than Sancho with comparable attributes values when Sancho is a good player. With Owen from Case 7, winning games for Stoke City, is still paid way less than Ronaldo who has not even been playing well. Such wage discrepancies are based on relative differences between the people rather than marginal output.

***From the discussion based on the above results, it can be said that the structuring of wages in top division leagues are more productive/performance based as the model could explain the wages of the players with an RMSE of 11064. But the model could not explain the same very well in the lower leagues as the wages were as low as 2000 with minor differences and variations in the skills of different players.***

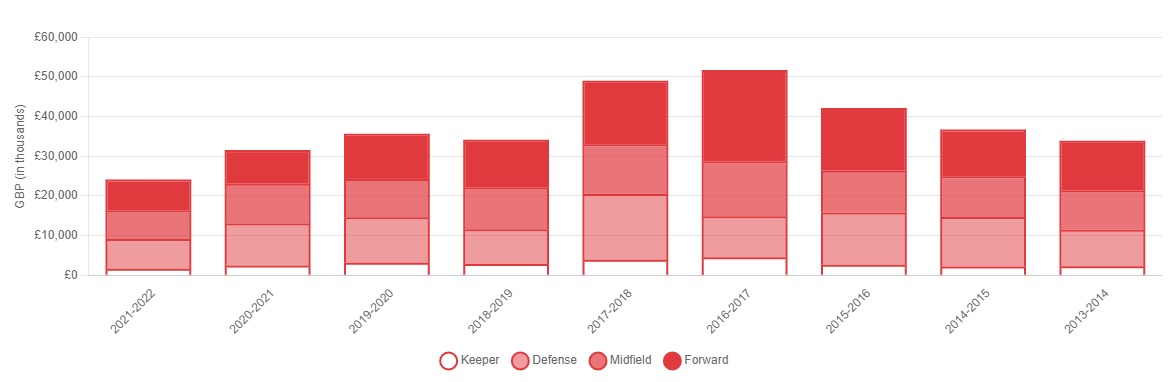
**5.2 PART B- CLUB REVENUE ANALYSIS**

Also, from the brief analysis on club revenues and wages, it is very clear that the top division leagues and its clubs have revenue support from various sources. Therefore, even in debt, the clubs keep performing well due to the constant in flow of cash. It is safe to say that this behaviour is common across at least four of the five leagues considered. Although, Italian clubs did not make the top 10, Juventus qualified for the 11th position with €400 million revenue for the 2021-22 season.



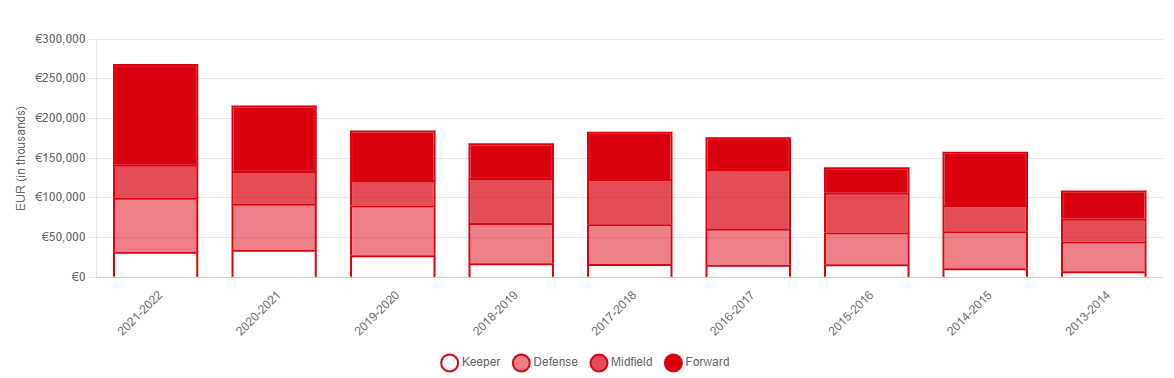
Figure 5.1 Club Revenue of Top 20 clubs

But the revenue and expenditure differ drastically amongst the clubs in the same country. This difference in the operations of the clubs is based on the leagues in which they play. Stoke City did fare well in the 2017-18 season when the club had the chance to play in the Premiere League. From the graph, SCFC’s deteriorating salaries scenario from 2017 season could be observed. On a general note, over the time, both the clubs tend to pay their forwards more than the other positions.



Graph 5.10 SCFC salary distribution based on player position over time

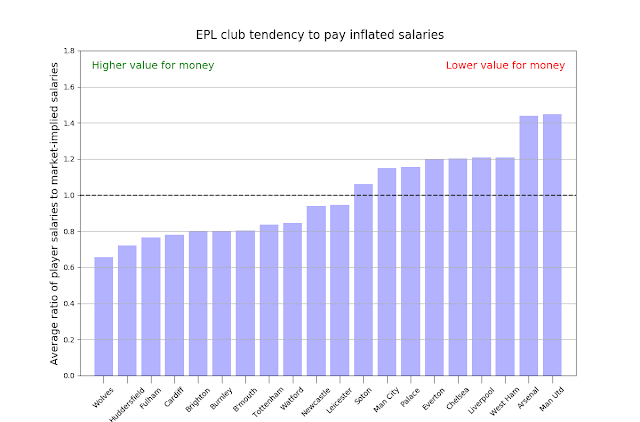
On the other hand, MUFC has had an increasing graph in terms of the salary fund. But it also pays ten times inflated salaries when compared to SCFC.



Graph 5.11 MUFC salary distribution based on player position over time

In fact, the ratio of transfer income to transfer expenditure calculated for SCFC and MUFC, in order to determine the amount of transfer income retained and spent, were 3 and 0.2 respectively which insinuates that the big clubs spend a lot on acquiring and sustaining their players as they have various sources of inflowing cash while the small clubs use these transfer fees from big clubs as an indispensable source of income.

It aligns with the notion that clubs who invest more in their players have more on-field performance. These clubs are mostly from the top division leagues across countries that tend to acquire the most talented players in the market and pay them more than their market rates to retain them. Also, a club's profitability is based on league place which can be seen from the amount of revenues garnered by the top clubs through various sources. Since small clubs have less talented players, sponsors sense more risk when it comes to investing on them and therefore, these clubs hardly make to the top divisions.

As per a study, the graph below represents the EPL clubs that are more prone to pay inflated salaries. Values above 1 signal that a club tends to pay salaries that are more than the market rate, while values below 1 suggest that they are lower (Eightyfivepoints, 2019).

Graph 5.12 EPL Clubs tendency to pay inflated salaries

The greatest offenders when it comes to overpaying their players are Manchester United and Arsenal, whose squads' average salaries are 40% higher than what their market value would indicate. Pogba, Sanchez, Mata, Young, Valencia, and Shaw are the six members of United's roster who make more money than their market-implied salaries. They have had a bad reputation for paying their players inflated salaries. With Manchester United’s disappointing performance in the season 2021-22, hiring an old but talented player like Ronaldo seemed like the last resort for the club to retain sponsors and its position in the EPL. Also, affording such an internationally reputed player was also not a problem for such clubs but the small clubs suffer. The fact that there is a huge discrepancy between what different clubs actually pay their players is unfair.

CHAPTER VI  
CONCLUSION

The findings demonstrate that professional athletes' pay are not simply determined at random, but rather are heavily influenced by a number of systematic characteristics, including age, experience, and performance, which are also common in other professions. The attributes those helped determine the wage were OVR, Value, International Reputation, Reaction and many more. This is largely consistent with findings of individual studies of Italy, Germany, etc. where players’ performances such as reaction and OVR affect their wages. But EPL additionally has an overpowering effect of International Reputation and Value as well in case of some players.

Sports teams' pay structures differ from those of regular occupations in that incomes are distributed more skewed overall and sports teams employ more demanding hiring criteria for positions. In more recent years, Matesanz et al. (2018) had proposed that European competitions, such as the UEFA Champions League or UEFA Europa League, are in fact a "money game" where the clubs that spend the most on transfers do better on the field. In five of the main European football leagues, Nsolo, et al. (2018) employed the technique to identify elite players and demonstrate that forwards do better in predictions than other positions.

In football, based on the above analyses, players reached their maximum salary and were highly valued between the age of 24 and 30. Additionally, certain players are more popular than others, irrespective of their age or performance, which significantly and positively influences their salary. Salaries in professional sports are generally right-skewed, with a relatively large number of players receiving low salaries and a few players earning far more. These large salary differences were explained by both ‘superstar’ and ‘tournament’ effects. Players not only desire to join more successful teams, but those teams also want to be identified with the former, and this influences wages and transfer fees significantly.

Football players rarely live up to their price tags. Some of the better examples of players who are worth the extra money include Cristiano Ronaldo, Zinedine Zidane, and most recently Virgil Van Dijk. It won't get better until football clubs stop overpaying for players and just pay what they're worth and what they could provide to the team, as opposed to paying for hype and attention. But my pessimistic side tells me it won't ever happen. By concentrating on the identification, growth, and ensuing profitable on-sale of players, numerous clubs have created a unique business model. Transfer fees have become into a crucial source of cash for some of these clubs especially small clubs.

EPL clubs are quite big spenders, acquiring seasoned players from other top leagues in Europe and beyond. But the tendency toward deregulation also gave rise to a number of abuses, such as the murky circumstances surrounding the operations of intermediaries or the invention of devices like third-party ownership in the transfer market. These factors all support the adoption of stricter regulations designed to guarantee the proper operation of the labour market for professional football players in order to prevent jeopardising the two primary attributes of professional sports: the fairness and the unpredictable nature of sporting events.

The income dominance of English clubs is not going to be challenged in the near future, but it makes one wonder how long it will be before all 20 Premier League clubs are represented in the top 30. Only the Premier League and La Liga have started new TV rights cycles for 2022–2023 among the top five European leagues. The demand from overseas broadcasters, which saw the value of international rights climb by €422 million each season (an increase of 26% for the 2022/23 to 2024/25 cycle compared to the 2019/20 to 2021/22 cycle), caused the Premier League to see an increase in the total value of its media rights. As rights agreements with current rights holders were "rolled over" for a further three seasons, the value of domestic media rights remained unchanged. La Liga, meanwhile, started a fresh domestic broadcast season. The league also sold several rights packages on a non-exclusive basis and/or controlled them directly, and it entered into agreements with broadcasters for longer five-year periods.

I don't think this issue can be solved easily. Yes, spending limits and procedures have been put in place by football associations all over the world to combat this, but none of them appear to be effective since every transfer window brings about a high-profile transfer that makes no sense at all. Despite the assumption of the study of the market being competitive especially after the Bosman Ruling, big clubs tend to monopolise the market in the absence of regulations.

This is detrimental to the sport right away. The larger and wealthier teams continue to pay inflated rates for players, leaving the smaller clubs without the resources they need to compete. Due to this, the league tables will have a significant disparity and the overall quality of football would drastically drop each season.

FUTURE WORK

The transfer market for women's football players is another crucial issue that requires additional attention. It is a new issue of labour market inequality that economists and regulators must fairly address. A detailed analysis on different types of football effects is also important but from the perspective of fans and media and how they perceive their starts.

The process for deciding transfer prices and wages is handled by the managements of the clubs and their agents, thus it is not entirely clear. To further regularise the market and provide a uniform method of determining and paying wages, the ambiguity must be eliminated. Deploying this model further on an online platform will help understand the importance of the model and make best use of it.

A systematic study where the academy players are vetted separately and then the allocation of funds to those young players can be compared across the leagues will help in better understanding the initiative that different clubs undertake to support the young footballers. Tracing their performances from early on in the dataset can aid in identifying their potential and the long term benefit for the respective clubs as well.

REFERENCES

Aha, D., Kibler, D and Albert, M.A. (1991). ‘*Instance-based learning algorithms,*' Machine Learning volume 6, pages37–66 (1991). Available at: https://link.springer.com/article/10.1007/BF00153759 [Accessed 20 October 2022].

Aldous, D. (1993). ‘*Approximate Counting via Markov Chains*', Statistical Science, Vol. 8, No. 1. Available at: http://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/ConvergeRates/Aldous.1993.pdf [Accessed 25 October 2022].

Allwright, S. (2022). ‘*What is a baseline model in machine learning?*. Available at: https://stephenallwright.com/baseline-machine-learning-models/ [Accessed 12 January 2023].

Amazon (2022). ‘*How XGBoost Works*.’ Available at: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html [Accessed 31 January 2023].

Arnedt, R. B. (1998). *European union law and football nationality restrictions: the economics and politics of the bosman decision. Emory International Law Review, 12, 1091*. Available at: https://heinonline.org/HOL/LandingPage?handle=hein.journals/emint12&div=27&id=&page= (Accessed: November 15, 2022).

Barrio, G., and Pujol, F. (2007). ‘*Hidden monopsony rents in winner-take-all markets-sport and economic contribution of Spanish soccer players*.' Managerial and Decision Economics, 2007, vol. 28, issue 1, 57-70. Available at: https://econpapers.repec.org/article/wlymgtdec/v\_3a28\_3ay\_3a2007\_3ai\_3a1\_3ap\_3a57-70.htm [Accessed 14 December 2022].

Brocard, J. and Lepitit, C. (2018). *The labour markets of professional football players,* [Online]. 1st Edition, 4-11. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9781351262804-24/labour-markets-professional-football-players-jean-fran%C3%A7ois-brocard-christophe-lepetit?context=ubx [Accessed 14 December 2022].

Brownlee, J. (2017). ‘*Why One-Hot Encode Data in Machine Learning?*.’ Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning [Accessed 23 January 2023].

Brownlee, J. (2020). ‘*How to Perform Feature Selection for Regression Data*.’ Available at: https://machinelearningmastery.com/feature-selection-for-regression-data/ [Accessed 23 January 2023].

Carmichael, F., & Thomas, D. (1993). *Bargaining in the transfer market: theory and evidence.* Applied Economics, 25(12), 1467-1476. Available at: https://www.tandfonline.com/doi/abs/10.1080/00036849300000150 (Accessed: November 19, 2022).

Carmichael, F., Forrest, D. and Simmons, R. (1999), ‘*The labour market in association football: who gets transferred and for how much?*’, Bulletin of Economic Research, 51 (2), 125–50. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8586.00075 (Accessed: November 19, 2022).

Carmichael, F. (2006). *The Player Transfer System*. In W. Andreff, & S. Szymanski, *Handbook of the Economics of Sports* (pp. 668 - 676). Cheltenham UK; Northampton MA, USA: Edward Elgar Available at: https://ideas.repec.org/h/elg/eechap/3274\_73.html (Accessed: November 19, 2022).

Deloitte (2023) “*Deloitte Football Money League 2023*”. 26th edition. Available at: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html [Accessed 31 January 2023].

Desai, S. (2022). RedBull. 2022. 8 fundamental skills you need to develop to become a better football player. [ONLINE] Available at: https://www.redbull.com/in-en/basic-football-skills-drills-training. [Accessed 4 November 2022]

Eightfivepoints (2019). ‘*From Sessegnon to Sanchez: How to calculate the correct market salary for EPL player’*. Available at: http://eightyfivepoints.blogspot.com/2019/05/from-sessegnon-to-sanchez-how-to.html (Accessed: November 24, 2022).

ERTHEO. (2022). *15 Key Soccer Skills – How to Achieve Success in Football* | Ertheo. [ONLINE] Available at: https://www.ertheo.com/blog/en/elements-success-in-football/. [Accessed 4 November 2022].

Ezzeddine, M. (2020) *Pricing football transfers : determinants, inflation, sustainability, and market impact : finance, economics, and machine learning approaches*, *ResearchGate*.net. ResearchGate. Available at: https://www.researchgate.net/publication/350174534\_Pricing\_football\_transfers\_determinants\_inflation\_sustainability\_and\_market\_impact\_finance\_economics\_and\_machine\_learning\_approaches (Accessed: November 15, 2022).

FIFA (2022). ‘*FIFA — SOCCER’S WORLD GOVERNING BODY*.’ Available at: https://www.ussoccer.com/history/organizational-structure/fifa [Accessed 30 November 2022].

Football History (2022). ‘*Manchester United FC*.’ Available at: https://www.footballhistory.org/club/manchester-united.html [Accessed 31 January 2023].

Football Stadiums (2022). *Football Player Transfers Explained.* Available at: https://www.football-stadiums.co.uk/articles/football-player-transfers-explained/ (Accessed: November 15, 2022).

Frick, B. (2006). *Salary determination and the pay-performance relationship in professional soccer: Evidence from germany*. Sports Economics After Fifty Years: Essays in Honour of Simon Rottenberg. Oviedo: Ediciones de la Universidad de Oviedo, 125–146. Available at: https://www.researchgate.net/publication/24131439\_Salary\_Determination\_in\_the\_German\_Bundesliga\_A\_Panel\_Study (Accessed: November 24, 2022).

Frick, B. (2007). *The football player’s labor market: Empirical evidence from the major european leagues. Scottish Journal of Political Economy*, 54(3), 422–446. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9485.2007.00423.x (Accessed: November 24, 2022).

Frick, B. (2008). *The impact of managerial quality on organizational performance: Evidence from German soccer,* Managerial and Decision Economics.Available at: https://www.researchgate.net/publication/46509524\_The\_impact\_of\_managerial\_quality\_on\_organizational\_performance\_Evidence\_from\_German\_soccer (Accessed: November 24, 2022).

Frick, B. (2011). *Performance, salaries, and contract length: empirical evidence from german soccer*. International Journal of Sport Finance, 6(2), 87. Available at: https://econpapers.repec.org/article/jsfintjsf/v\_3a6\_3ay\_3a2011\_3ai\_3a2\_3ap\_3a87-118.htm (Accessed: November 24, 2022).

Friedman, A.L. (2002). ‘*Developing Stakeholder Theory,*' Journal of Management Studies 39(1):1-21. Available at: https://www.researchgate.net/publication/227375374\_Developing\_Stakeholder\_Theory [Accessed 25 October 2022].

GeeksforGeeks (2022). ‘*Decision Trees*.’ Available at: https://www.geeksforgeeks.org/decision-tree [Accessed 31 January 2023].

Goal (2022). ‘*Manchester United*.’ Available at: https://www.goal.com/en-gb/team/manchester-united/6eqit8ye8aomdsrrq0hk3v7gh [Accessed 31 January 2023].

Hausman,J. A. and Leonard, G. K. (1997). *'Superstars in the National Basketball Association: economic value and policy.*' Journal of LaborEconomics, vol. 15 (4) (October), pp. 586-624. Available at: http://dx.doi.org/10.1086/209839 [Accessed 14 December 2022].

Hessayon, A. (2014) *From violent peasants to multi-million pound megastars: The History of Football*, *Academia.edu*. Available at: https://www.academia.edu/7307120/From\_violent\_peasants\_to\_multi\_million\_pound\_megastars\_the\_history\_of\_football (Accessed: November 15, 2022).

http://eightyfivepoints.blogspot.com/2019/05/from-sessegnon-to-sanchez-how-to.html

https://www.transfermarkt.com/

https://www.capology.com/

https://sofifa.com/

IBM (2022). ‘*What is Linear Regression’*. Available at: https://www.ibm.com/uk-en/topics/linear-regression [Accessed 31 January 2023].

Iterpro (2022). How do football clubs make money,. Available at https://iterpro.com/how-do-football-clubs-make-money/ (Accessed: November 25, 2022).

James, S. (2022). “*How do you value a player?”,*The Athletic [ONLINE] Available at: https://theathletic.com/3085749/2022/01/27/premier-league-how-do-you-value-a-player/. [Accessed 9 November 2022].

Kumar, D. (2018). ‘*Introduction to Data Preprocessing in Machine Learning* ', Beginners Guide for Data Preprocessing. Available at: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d [Accessed 2 January 2023].

Kohavi, R. (1995). ‘*A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. Available at: https://www.researchgate.net/publication/2352264\_A\_Study\_of\_Cross-Validation\_and\_Bootstrap\_for\_Accuracy\_Estimation\_and\_Model\_Selection [Accessed 25 October 2022].

Rosen, S. and Sanderson, A. (2001*) Labour Markets in Professional Sport.s* Available at: https://www.jstor.org/stable/2667957 (Accessed: November 15, 2022).

Roy, B. (2020). ‘*All about Feature Scaling*.’ Available at: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35 [Accessed 23 January 2023].

Ruijg, J., and Ophem, H. (2014). ‘*Determinants of football transfers*'. Available at: https://econpapers.repec.org/paper/amewpaper/1401.htm [Accessed 25 October 2022].

Saha, P. (2019). ‘*Exploratory Data Analysis — Unravelling a story with data*'. Available at: https://towardsdatascience.com/exploratory-data-analysis-unravelling-a-story-with-data-b01d70069035 [Accessed 2 January 2023].

Saini, A. (2021). ‘*AdaBoost Algorithm – A Complete Guide for Beginners’*. Available at: https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/ [Accessed 21 January 2023].

Shin, J., and Gasparyan, R. (2014). ‘*A novel way to Soccer Match Prediction*'. Available at: http://cs229.stanford.edu/proj2014/Jongho%20Shin,%20Robert%20Gasparyan,%20A%20novel%20way%20to%20Soccer%20Match%20Prediction.pdf [Accessed 27 October 2022].

Sloane, P.J. (1971). ‘*The Economics of Professional Football: The Football Club as a Utility Maximiser*.' Scottish Journal of Political Economy 18(2):121-46. Available at: https://www.researchgate.net/publication/4789346\_The\_Economics\_of\_Professional\_Football\_The\_Football\_Club\_as\_a\_Utility\_Maximiser [Accessed 14 December 2022].

Sohns, J. (2021). ‘*fifa-ratings-explained-overall-rating’*. Available at: https://earlygame.com/fifa/fifa-ratings-explained-overall-rating [Accessed 31 January 2023].

Speight, A., and Thomas, D. (1997). ‘*Football league transfers: a comparison of negotiated fees with arbitration settlements*', Applied Economics Letters. Available at: https://www.tandfonline.com/doi/abs/10.1080/758521830 [Accessed 27 October 2022].

Szymanski, S., and Smith, R. (1997). *'The English Football Industry: Profit, Performance and Industrial Structure.*' International Review of Applied Economics 11(1):135-153. Available at: https://www.researchgate.net/publication/227351647\_The\_English\_Football\_Industry\_Profit\_Performance\_and\_Industrial\_Structure [Accessed 14 December 2022].

Szymanski, S., and Kuypers, User. (1999). ‘*A Market Test for Discrimination in the English Professional Soccer Leagues* ', Journal of political Economy, 2000. Available at: https://www.journals.uchicago.edu/doi/abs/10.1086/262130 [Accessed 25 October 2022].

Szymanski, S. (2015). ‘*Making Money Out of Football* ', Scottish Journal of Political Economy 62(1). Available at: https://www.researchgate.net/publication/271021120\_Making\_Money\_Out\_of\_Football [Accessed 25 October 2022].

Transfer Markt. 2022. *Football history*. [ONLINE] Available at: https://www.transfermarkt.co.uk/. [Accessed 9 December 2022].

Vrooman, J. (1996). ‘*The baseball players' labor market reconsidered*', Southern Economic Journal. Available at: https://scholar.google.com/citations?view\_op=view\_citation&hl=en&user=jgQTYFIAAAAJ&citation\_for\_view=jgQTYFIAAAAJ:9yKSN-GCB0IC [Accessed 25 October 2022].

Wood, R. (2008). RedBull. 2022. *Soccer History Timeline*, Topend Sports Website, 2008, [ONLINE] Available at: https://www.topendsports.com/sport/soccer/history.htm. [Accessed 5 February 2023].

Wu, S. (2020). ‘*Multicollinearity in Regression*.’ Available at: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea [Accessed 23 January 2023].

Yadav, A. (2018). ‘*SUPPORT VECTOR MACHINES(SVM)*. Available at: https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589 [Accessed 21 January 2023].

Yaldo, L., and Shamir, L. (2017). ‘*Computational Estimation of Football Player Wages,*' International Journal of Computer Science in Sport 16(1). Available at: https://www.researchgate.net/publication/318655683\_Computational\_Estimation\_of\_Football\_Player\_Wages [Accessed 20 September 2022].

Yiu, T. (2019). ‘*AdaBoost Algorithm – A Complete Guide for Beginners’*. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed 21 January 2023].

APPENDIX A- MULTICOLLINEARITY

Based on heatmap’s correlation results, following variables were removed to avoid multicollinearity

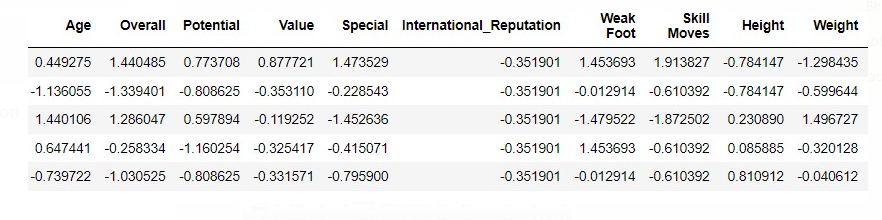


APPENDIX B - FEATURE SCALING

Before Feature scaling, the raw data as it appeared to be is shown below



After feature scaling, the standardized data as it appeared in notebook



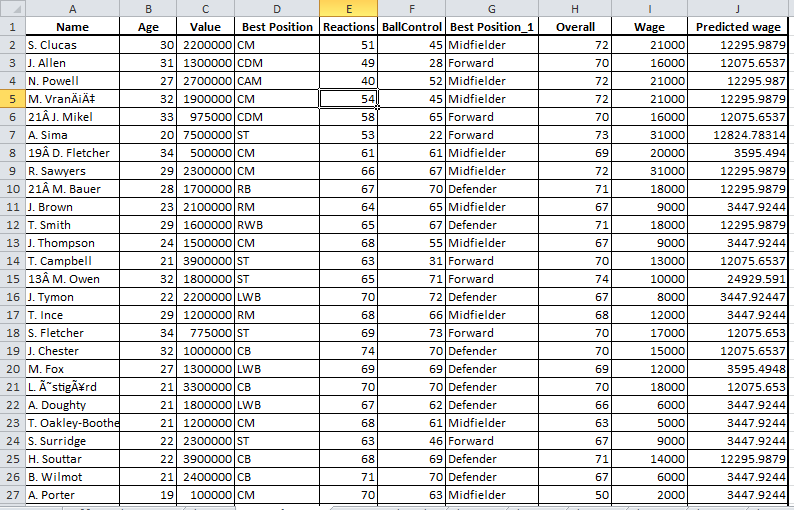
APPENDIX C- CLUB REVENUE

Data for club revenue and its constituents along with other sources of income and expenditure for the top 10 clubs of season 2021-22 has been given below



APPENDIX D- SCFC

The dataset for Stoke City Football Club with monotonic variables (feature importance) has been shown below. Both actual and Predicted wages have been shown.



APPENDIX E- MUFC

The dataset for Manchester United Football Club with monotonic variables (feature importance) has been shown below. Both actual and Predicted wages have been shown.

