University of Cape Town

Masters Dissertation

An affordable data solution for player recruitment for clubs in the South African Premier Soccer League

Author:
Wesley King
Supervisor:
Neil Watson

Student Number:
KNGWES002

June 25, 2025

List of Figures

List of Tables

Contents

1 Abstract

As football becomes increasingly data-driven, the high cost of advanced player analytics threatens to leave resource-limited clubs at a competitive disadvantage, particularly in player scouting. This growing reliance on expensive, granular data underscores the need for affordable, innovative data solutions. This dissertation seeks to democratize access to player evaluation data for football clubs in the South African Premier Soccer League. This is achieved by developing a cost-effective system that uses models to approximate Statsbomb's proprietary 'On the ball' player evaluation metric using cheaper, frequency data from Wyscout and FBref. The analysis shows that linear regression models can effectively estimate key components of this metric using basic frequency statistics. The findings are then packaged into a prototype web-based Decision Support System with budget-aware scouting features, showcasing how club scouts and analysts can integrate sophisticated data-driven recruitment strategies into their clubs without incurring prohibitive data costs.

2 Introduction

2.1 Contextualization

Football is one of the most lucrative and globally influential industries of the modern era (Yiapanas et al., 2024). With billions of pounds circulating through player transfers, sponsorship deals, and broadcasting rights, the sport has evolved into an economic powerhouse (Aygün et al., 2023). In recent years, the introduction of spatiotemporal tracking data has driven a data-based revolution as teams aim to utilize advanced analytics involving areas like artificial intelligence and statistical modeling to gain a competitive advantage (Link et al., 2016; Forcher et al., 2022; De Silva et al., 2018; Azmat et al., 2024).

Recent advancements in analytical techniques applied to football have significantly altered both gameplay strategies and the management and organization of clubs and teams (Thakkar & Shah, 2021). A particular focus on player evaluation techniques has seen scouting innovations improve a club's ability to sign undervalued players who have been overlooked by traditional scouting practices, such as subjective eye-witness reports (Lawlor et al., 2021). However, the acquisition of the required data for these practices can entail substantial costs, making it inaccessible to many teams. This reveals a stark reality: the unequal distribution of wealth among football clubs worldwide (Szymanski, 2001).

While comparatively few teams enjoy the riches of television revenues and corporate sponsorships, many others struggle to compete financially, languishing beneath their wealthier counterparts (Morrow, 2023). This enables economically advantaged clubs to leverage the latest technologies (Bertheussen, 2023). Such technologies can enhance various aspects of a club's operations, including player recruitment, thereby widening the performance gap between clubs with differing financial resources.

Considering this inequality, there arises a pressing need for a simple, affordable approach to football scouting and analytics. Recognizing the transformative potential of data-driven methodologies in particular, this study demonstrates the development of a scouting system that uses statistical models to predict the player evaluation metric On-Ball-Value (OBV) (Hudl StatsBomb, 2021a), created by data providers Statsbomb (Hudl StatsBomb, 2024), using more accessible, cost-effective data sources. This system aims to accommodate and improve teams with smaller budgets by providing an elite scouting system at a fraction of the usual cost. An accompanying Decision Support System (DSS) has also been developed that operationalizes these methodologies, enabling teams to effectively identify potential transfer targets. The Premier Soccer League (PSL) in South Africa, which contains clubs with limited transfer budgets as evidenced by modest player acquisitions

compared to international leagues (e.g., England and France) (Transfermarkt, n.d.), presents an ideal environment for implementing cost-effective recruitment strategies. As a result, the approach developed in this study has the potential to democratize access to data analytics within the footballing community, empowering clubs of all financial standings to establish data-driven scouting departments.

2.2 Overview of Football and Data (in Football)

Football, also known as soccer in some regions, is characterized by two teams of eleven players competing on a designated field. Each team fields ten outfield players and one goalkeeper, with the latter uniquely permitted to handle the ball within their designated area. This area, commonly referred to as "the box," is demarcated by the rectangular zones illustrated on either end of Figure 1. The shape of a pitch is also shown in Figure 1 below, and is generally 105m long by 68m wide, as recommended by FIFA (FIFA, n.d.).

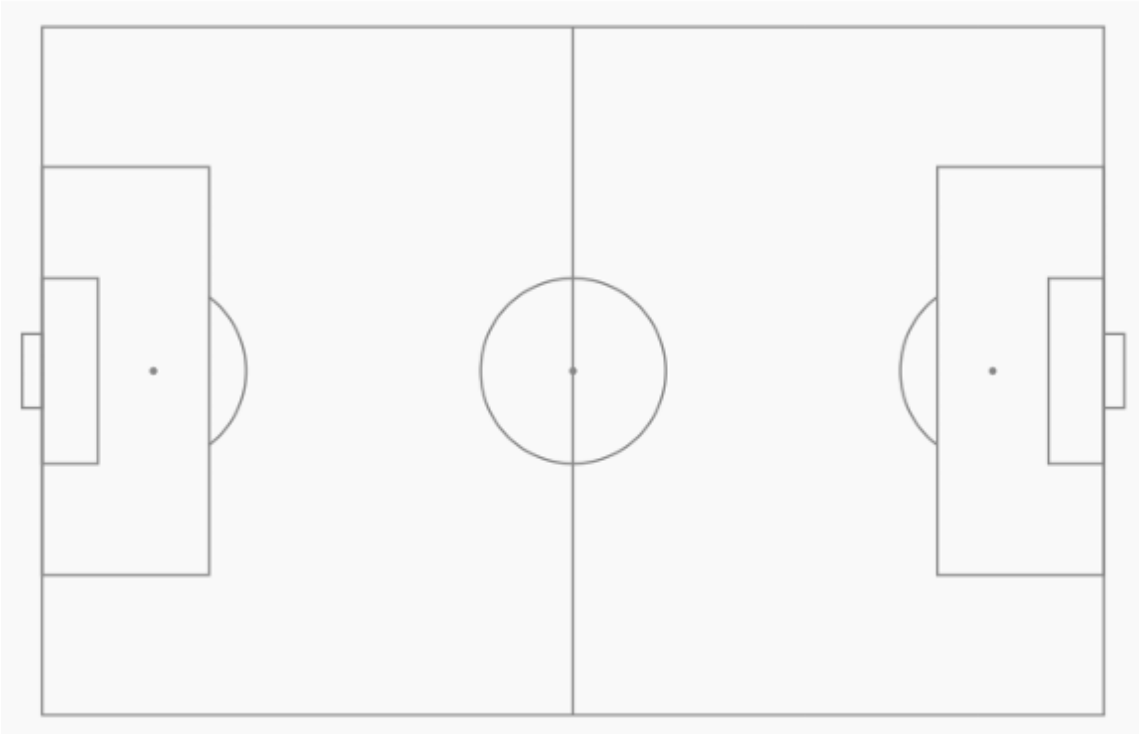

Figure 1: A football field

A football game is comprised of two halves of approximately forty-five minutes each, during which each team attempts to kick or head the ball into the opposition team's goals. A goal increases the score for the team who scored, and at the end of the

two halves the team with the most goals wins the game, or if the scores are even the match is deemed a 'draw'. Figure 2 below shows the general playing positions in football: Goalkeeper, Defender, Midfielder and Attacker. The Goalkeeper is the last line of defense. Their primary role is to prevent the opposing team from scoring by guarding the goal and commanding the defensive line. In the modern game, goalkeepers also initiate plays from the back by distributing the ball to teammates to start an attack (Otte et al., 2022). A Defender is generally responsible for stopping the opposing team from creating scoring opportunities. Defenders can be further divided into center-backs, who focus on protecting the central area in front of the goal, and fullbacks, who guard the wide areas and support the attack by delivering crosses into the opposition's penalty area (Zhang et al., 2024). Thus, center backs are a lot more defensively minded, while fullbacks, who are also defenders, play more attacking football. A Midfielder acts as a link between defense and attack, playing a crucial role in both breaking up the opponent's plays and creating scoring opportunities (Aalbers & Haaren, 2018). This sees midfielders have more specialized roles such as defensive midfielders, who primarily focus on protecting the defense, and attacking midfielders, who are more involved in creating and scoring goals. One also gets wide midfielders, known as wingers, who, like the fullbacks, operate on the flanks and deliver crosses into the box. A forward's primary role is to score goals and create scoring opportunities.

A critical difference between football and other sports like Basketball or American football is that it is a relatively low-scoring game, with an average of 1.4 goals being scored per game in both women's and men's football (Sumpter, 2016). Thus, any Player Evaluation (PE) metric which captures a holistic view of a player's performance will be centered around valuing players by their actions when most often these actions did not contribute to scoring a goal. Furthermore, given that most of the actions in a game do not lead to goals within the same possession sequence, to solely rely on frequency measures such as goals and assists (traditional metrics) would be to ignore most of a player's game. This is where traditional metrics fail to represent a holistic measure of a player's quality, while data-driven models do well, as they can capture the performance of players throughout the game, considering each touch, where it occurs on the pitch, and its subsequent impact. This allows player actions to be valued in numerous different ways, whether it be valuing passes based on where on the pitch they occur, or valuing tackles based on the predicted attacking play that was prevented by the defensive action. This deepens the analysis beyond simpler frequency statistics such as number of progressive passes and number of tackles and facilitates a more nuanced perspective on a player's game.

Research in football analytics broadly divides into two primary categories. One stream focuses on match outcome prediction (Berrar et al., 2019), while the other

Figure 2: General positions on a football pitch (Modric et al., 2020)

concentrates on quantifying and evaluating specific in-game actions (Dick & Brefeld, 2022). Fully utilizing player evaluation metrics demands significant financial investment, not accessible to many PSL clubs. An unexplored field in football evaluations is the development of cost-effective player evaluation metrics using cheaper, more accessible frequency data. By allowing for local and foreign player markets to be accurately explored for a fraction of the cost, the gap in scouting capabilities between wealthier clubs and poorer clubs can be narrowed.

Central to this study is bypassing the need to rely solely on expensive data procurement from established providers, like Statsbomb, which charges approximately R200 000 for a single season of league data (Hudl StatsBomb, 2024). Instead, clubs with limited resources can adopt a more cost-effective approach by utilizing alternatives like Wyscout (Wyscout, 2024), and FBref (FBref, 2025). For the same price as buying a single league from Statsbomb, Wyscout provides data from over 50 leagues worldwide—an investment representing less than 1% of a PSL club's annual allocation-while FBref provide frequency statistics for 5 of the top leagues in Europe (England, France, Italy, Spain and Germany) for free. By leveraging predictive modeling to enhance data insights across Wyscout's broad league coverage, clubs can achieve comprehensive analytical capabilities at a fraction of the cost. Additionally, this approach serves as an entry point for larger European teams to integrate data analytics into their scouting departments utilizing FBref data, offering a budget-friendly avenue to get started with implementing important data-driven player performance information.

2.3 Objectives of Research

This study proposes a solution to address the disparity in scouting methods and opportunities between wealthy and less-wealthy clubs. By modeling Statsbomb's player evaluation metric *OBV* using more affordable frequency-based data from Wyscout (Wyscout, 2024) and free data from FBref (FBref, 2025), we aim to create a tiered scouting system which allows teams to affordably produce scouting reports for leagues worldwide. Although the primary focus is on the PSL, the findings of this dissertation will also allow teams in Europe's top 5 leagues (the leagues that FBref covers, such as first divisions in England, Spain, France, Germany, and Italy) to affordably incorporate data analysis into their player recruitment strategies, without needing to incur associated large upfront data acquisition costs. To achieve the above objective, this dissertation explores the application of various statistical and machine learning models to determine the most effective approach for capturing the complexity of *OBV* using simpler, more accessible frequency variables. This study also details the development of a prototype DSS that visualizes the results, demon-

strating how the methodology developed in this study can be applied in practice to leverage data-driven player performance information to support player scouting decisions.

3 Literature Review*

3.1 Data-Driven Success: The Brentford and Brighton Revolution

The transformative impact of data analytics in football is exemplified by clubs such as Brentford F.C. and Brighton & Hove Albion F.C. Inspired by the Moneyball approach (Triady & Utami, 2015), Brentford's recruitment strategy has focused on identifying undervalued players through advanced analytics. Brentford's owner, leveraging his expertise in sports betting and data analysis, instilled a data-driven ethos within the club (Wigmore, 2017). This approach has led to remarkable financial success, with Brentford recording a net transfer profit of £108.82 million over nine years, enabling investments in infrastructure like their state-of-the-art stadium. By focusing on underlying performance metrics instead of short-term results, Brentford has ensured stability and sustained progress, even amidst managerial changes (O'Brien, 2020). Their commitment to metrics over match results has fostered a long-term vision, allowing them to make difficult but necessary decisions for the club's future.

Brentford's success underscores the versatility of data analytics beyond recruitment, as it is used to evaluate team performance and predict future trends. The club prioritizes underlying metrics such as expected goals and passing efficiency over league standings, helping to mitigate the impact of luck in low-scoring games (O'Brien, 2020). For instance, decisions regarding managerial appointments and dismissals are informed by a nuanced understanding of these metrics, ensuring that overachievement or underachievement is not merely attributed to chance. By exploiting inefficiencies in player markets and emphasizing a data-centric evaluation of performance, Brentford demonstrates the critical role data plays in modern football, particularly for financially modest clubs competing in the Premier League.

Similarly, Brighton FC has embraced data analytics to transform the club from a struggling third-tier side to a competitive Premier League team. Brighton's owner also has a history in sports gambling, as he is the founder of betting company Starlizard. He has thus managed to apply his data expertise to revolutionize recruitment at Brighton (Garratt-Stanley, 2023a). The club focuses on scouting underutilized markets, such as leagues in South America and East Asia, identifying undervalued players with high potential. This strategy has yielded significant profits, with recent sales including £76 million for an Ecuadorian midfielder and £47 million for a Spanish left-back, allowing Brighton to reinvest in their scouting process. Brighton's

[0]Certain portions of this section have been adapted from my honours dissertation (King & Bhorat, 2022)

ability to replace departing players seamlessly, often at a fraction of the cost, highlights their efficient use of data.

Brighton's reliance on data has also extended beyond recruitment to tactical and managerial decision-making too. Their manager noted the initial adjustment required to adapt to Brighton's data-driven approach, which integrated algorithms with video analysis to provide deeper insights into player and team performance (Garratt-Stanley, 2023b). This multifaceted approach contrasts with traditional reliance solely on video footage, emphasizing the value of analytics in modern football (Naylor et al., 2023). Despite significant stylistic differences between Brighton and Brentford (McDonnell & Sisneros, 2023), both clubs share a foundational reliance on data, demonstrating its adaptability to various playing styles.

Brentford F.C. and Brighton Hove Albion F.C. demonstrate how data-driven approaches can transform football operations, particularly in recruitment and performance analysis. Their ability to consistently compete with and outperform wealthier clubs (McDonnell & Sisneros, 2023) provides a compelling blueprint for PSL clubs seeking to maximize their resources. Their success illustrates that strategic implementation of data analytics can help bridge financial gaps in football, offering a viable pathway for South African clubs to enhance their competitiveness despite resource constraints.

3.2 Economic Context of the Premier Soccer League (PSL)

Much like the Premier League in England, the Premier Soccer League (PSL), has long been dominated by financially powerful clubs. These clubs include Kaizer Chiefs, Orlando Pirates, and Mamelodi Sundowns. These teams leverage expansive fan bases and lucrative sponsorship agreements—exemplified by high-profile partnerships with Vodacom—to secure transfer budgets that far exceed those of their less affluent counterparts. In contrast, less wealthy clubs must primarily rely on the standard PSL monthly allocation of 2 million rand. This financial muscle possessed by clubs such as Mamelodi Sundowns has enabled them to assert their dominance, as evidenced by their recent capture of six consecutive PSL titles and a net transfer expenditure exceeding eight million euros over two seasons. Such disparities create a challenging competitive environment for smaller clubs, which, constrained by tighter budgets (Transfermarkt, n.d.), are unable to compete solely through financial means. This concentration of financial power amongst these 3 teams reinforces their competitive edge, whilst ensuring less affluent clubs have to explore innovative strategies in order to compete.

In a sport where success has historically been closely correlated with financial supremacy (Rohde & Breuer, 2016), the emergence of data-driven approaches—particularly

in recruitment—offers opportunities for clubs to outmaneuver rivals without relying solely on monetary resources (Herberger & Litke, 2021a). As demonstrated by Brighton and Brentford in the Premier League, the integration of advanced analytics into recruitment processes can provide a pathway for under-resourced teams to compete more effectively. This shift presents a promising avenue for smaller PSL teams to challenge their wealthier domestic rivals through strategic recruitment intelligence, while simultaneously enabling European clubs to enhance their reputations without being solely dependent on funding.

3.3 Player evaluation (PE) metrics

The football industry serves as an area of interest amongst scholars in the academic domain. This interest is reflected in the growing amount of football analytics studies, which look deeper into valuing players utilising structured data as oppose to non-statistical video analyses. These efforts in developing player evaluation (PE) metrics aim to better quantify a player's footballing abilities using objective measures (Herberger & Litke, 2021b). Below lies a critical examination of the evolution of these PE metrics, before the current state-of-the-art is arrived at, which is the PE metric modeled in this dissertation(OBV), using more affordable Wyscout data.

3.3.1 Evolution of PE metrics

Before companies such as Statsbomb, Statsperform (Opta, 2025) and Skillcorner (SkillCorner, 2024) became household names in the football industry, bringing innovation to PE by utilising Markov Chains and deep machine learning models, simpler frequency-based measures were standard in football performance analysis. One of the earliest attempts to quantify a player's influence on a match was the Flow Centrality score developed by Duch et al. (2010). This metric measured the proportion of times a player was involved in a sequence of plays leading to a shot. However, as the authors noted, this approach had significant limitations, particularly in its tendency to overvalue midfielders and attackers while undervaluing defenders, who are less frequently involved in build-up play. Building on this, McHale and Relton (2018) sought to identify key players within teams by employing network analysis, which examined player connections and aggregated pass difficulty. While this provided valuable insights, defenders and goalkeepers were again found to score lower than midfielders and attackers, as their passes were generally less challenging. Further research focused specifically on valuing passes included the work done by Chawla et al. (2017), who leveraged spatiotemporal data to classify passes as good, bad, or ok. Expanding on this, Power et al. (2017) developed models which predicted both the risk and reward of passes, incorporating both factors to assess pass value more comprehensively. Meanwhile, Bransen et al. (2019) created a novel measure designed

to capture a pass's value in setting up scoring opportunities, further advancing the methodology for evaluating passing contributions.

Like frequency-based measures, the Top-down technique is another PE tool that isn't used as much anymore. This approach involved first assessing the team's overall effectiveness and then distributing credit proportionally among the team. The primary framework for this type of evaluation is the Plus-Minus (PM) rating system, which seeks to quantify a player's impact by comparing the team's performance with and without their presence. One of the earliest implementations of a PM-based assessment in football was introduced by Hvattum and Sæbø (2015), where matches were divided into discrete intervals (segmented by substitutions or dismissals), and players were assessed based on the goal differential during those periods, adjusted according to the duration of the segment. This methodology was later refined by Sæbø and Hvattum (2018), as they incorporated an additional factor to account for the competitive strength of the leagues in which the teams participated. Building upon the foundations laid by Hvattum and Sæbø, Kharrat et al. (2020) extended the PM regression framework by integrating expected goals (xG) as a key variable and introducing an alternative model where the variation in xG served as the primary explanatory factor. Departing from traditional regression-based approaches in PM ratings, Schultze and Wellbrock (2017) proposed a modified PM model that factored in both the significance of a goal and the relative caliber of the opposition. However, as noted by Hvattum (2019), this methodology introduced additional complexity, making cross-team player comparisons more challenging.Warnke and Sittl (2016) further refined Sæbø and Hvattum's PM system by adopting fixed-effects models, decomposing team success into distinct contributions from individual players, coaches, and overall squad strength, each of which captured short-term, mid-term, and long-term influences, respectively. In another innovative extension, Matano et al. (2018) implemented a Bayesian modeling framework, leveraging prior player valuations based on ratings from the widely recognized video game FIFA. Their study demonstrated that this Bayesian approach outperformed conventional multiple linear regression techniques in predicting match outcomes within basic Plus-Minus structures. Additionally, the Elo rating system, originally developed for chess, has been adapted across various sports. This system calculated a team or player's relative skill level by comparing expected versus actual performance outcomes, where ratings were adjusted after each match where winners gain points from losers based on the pre-match probability of victory (Wolf et al., 2020).

Early player evaluation systems took a bottom-up approach, calculating player scores by assigning weights to different actions and summing their values. The EA Sports Player Performance Index (PPI), introduced in 2004 (McHale et al., 2012), exemplified this methodology. It evaluated players using six subindices, including goals,

assists, and general match contributions. However, this system had a significant limitation: it didn't incorporate spatiotemporal data. Since GPS tracking wasn't widely implemented in English Leagues when the PPI was developed, the system couldn't account for the location where actions occurred. This led to an oversimplified valuation system where similar actions were assigned equal importance regardless of their spatial context. Later research by Klemp et al. (2021) explored more sophisticated approaches to modeling in-game dynamics, specifically focusing on predicting match outcomes in real-time. Their findings revealed that despite applying advanced feature engineering to event tracking data, in-game statistics provided surprisingly little predictive value compared to pre-game metrics when forecasting goals. Nevertheless, they acknowledged that event data remained valuable for understanding and explaining game dynamics.

In 2016, Brooks et al. (2016) developed the Pass Shot Value (PSV) metric, which evaluated passes based on their relationship to shot generation and their locations on the field. While this metric provided valuable insights, it showed a notable bias toward midfielders and attackers, potentially undervaluing defenders who excelled in defensive actions like tackles and interceptions rather than progressive passing. The metric's narrow focus also meant it overlooked other crucial aspects of the game, such as dribbling ability and goalkeeper saves.

Gyarmati and Stanojevic (2016) advanced the field with their QPass metric, which introduced a more sophisticated approach using team-specific grid systems overlaid on the pitch. Each grid location was assigned a value based on its correlation with goal-scoring opportunities relative to opponent chances. Pass values were calculated by measuring the change in probability between the starting and ending locations of the ball, effectively quantifying how each pass either increased or decreased shooting opportunities compared to the opposition. This groundbreaking work pioneered the use of spatial grids for differentially valuing possession based on pitch location. However, the metric had notable limitations: forwards typically received lower ratings due to already occupying dangerous positions, while goalkeepers were disproportionately favored since their passes naturally moved the ball toward the opposition goal. Like its predecessors, QPass's exclusive focus on passing prevented it from serving as a comprehensive player evaluation tool.

Decroos et al. (2018) developed a groundbreaking approach to action valuation with their Valuing Actions by Estimating Probabilities (VAEP) framework. This system evaluated actions based on their impact on the probabilities of scoring and conceding in subsequent phases of play. VAEP incorporated multiple contextual factors: the sequence of events leading to an action, the action's location, and the ball's position after completion. By leveraging event data, VAEP achieved more precise location-aware evaluations. However, the framework had notable limitations: it focused

exclusively on successful actions and didn't account for defensive contributions like tackles and interceptions. Building on this foundation, Singh (2020) introduced the "expected threat" (xT) model, which calculated the probability of scoring within k subsequent actions from any given pitch location. The model considered both immediate shooting opportunities and the potential for movement leading to shots. xT utilized a fixed grid generated by a Markov Model, computing average scoring probabilities from different locations across a league season. Like VAEP, xT leveraged event data to incorporate spatial information and provided a framework for valuing successful ball progression through passes and dribbles. However, it shared similar limitations in not accounting for shots, defensive actions, or player mistakes. Expected goals (xG) emerged as another significant metric for evaluating attacking efficiency (Green, 2012). This measure calculates scoring probability based on shot location, specifically considering the distance and angle to goal. While xG effectively quantifies finishing ability, it falls short as a comprehensive measure of attacking quality. Players who excel in playmaking and contribute significantly to attacking build-up may not receive appropriate recognition through this metric, highlighting a key limitation of xG in player evaluation.

Opta, a global sports data company, released and then updated their first player evaluation tool called 'Possession Value' (PV) (Stats Perform, 2024) throughout the 2019/20 season. This was implemented using client and user feedback. Departing from their original framework, which assessed the probability of a team scoring from their current possession, the updated PV framework adopts a time-based approach. Specifically, this measures the probability of the team in possession scoring within the subsequent 10 seconds. This strategic shift was informed by its demonstrably superior model performance when compared to alternative methodologies (Whitmore, 2020). To illustrate, envision Manchester City are in possession, specifically, Player A is in possession near the halfway line, where the PV stands at 1% ($PV_{start} = 0.01$). This means that the probability of Manchester City scoring from this possession chain within the next 10 seconds is 1%. Should he progress down the line and deliver a pass to Player B, who is situated within the box, this would elevate the possession value to 13% ($PV_end = 0.13$). The interpretation is that Player A effectively increases his team's likelihood of scoring within the next 10 seconds by 12%. Thus, PV acknowledges Player A's contribution irrespective of Player B's subsequent actions with the ball. In this way, PV isolates and values players' actions and their effect on the team's immediate chances of scoring. The model is hampered however by being trained on goals scored, and not xG, which has been proved to be a better predictor of future goals scored (octosport.io, 2022). The model would also be trained on a larger dataset if all shots were considered, as opposed to only ones that lead to goals.

While traditional player evaluation focuses on what happened during matches, counterfactual analysis techniques allow us to explore hypothetical what-if scenarios to better understand a player's decision-making and potential impact. Counterfactual evaluations represent an innovative approach to action valuation by analyzing hypothetical sequences of play that could have occurred under different circumstances. Van Roy et al. (2021) demonstrated this approach using Markov Decision Processes (MDP) to analyze optimal shooting locations for teams. Their research evaluated both immediate shooting opportunities and the potential value of maintaining possession for additional actions. By modeling team-specific attacking sequences through MDP, they generated optimized shooting policies and demonstrated that teams were consistently scoring below their potential. Their analysis culminated in the creation of team-specific shooting grids that mapped optimal scoring locations. In the defensive domain, Merhej et al. (2021) extended Singh (2020)'s work by developing the DaXt model to evaluate defensive actions. This model assessed tackles and interceptions by simulating the likely sequence of events that would have unfolded without the defensive intervention. The value of each defensive action was calculated by applying an xT framework to these hypothetical prevented sequences. While this approach innovatively quantified defensive contributions, it had two significant limitations: it couldn't provide comprehensive player evaluations, and it exhibited a frequency bias. This bias manifested in defenders from lower-ranked teams receiving consistently higher ratings than those from top teams, primarily because they had more opportunities to make defensive actions due to facing more attacking pressure.

The limitations of the player evaluation methods discussed above are clear. Whether focusing on isolated actions that fail to provide a comprehensive evaluation tool, providing values which are biased towards attacking players, or not considering penalizing players for unsuccessful actions, these metrics all have areas where they can be improved.

3.3.2 On-Ball Value (OBV)

OBV (Hudl StatsBomb, 2021a) is a quantitative measure that can describe the quality of football players irrespective of their position and will serve as the basis by which players in this dissertation are evaluated. *OBV* allows players' contributions to a team performance to be viewed on the same scale, allowing, for example, the contribution of a defender to be directly compared to the contribution made by a midfielder. This is one of the first possession value models which attempts such a feat, as models like xT and VAEP struggle to do this since they do not account for all actions (Singh, 2020), while Statsperform's PV metric has its own shortcomings as discussed earlier (Whitmore, 2020).

The *OBV* framework is a proprietary 'blackbox' model which evaluates each action by quantifying a proxy for the change in the team's probability of scoring and conceding attributable to that action. The model's approach is thus dual in nature, with two primary components quantified per action: 1) the change in the likelihood of conceding, and 2) the change in the likelihood of scoring. Some actions have high *OBV* for one component, and low for the other, with each action's *Total OBV* being the sum of these two quantities. Figure 3 below shows this. Plot A shows an action that largely decreases the probability of conceding, while only slightly increasing the probability of scoring. An action expected to achieve such a value would be a tackle in one's own defensive area, as this doesn't mean the defending team is much more likely to score, but it does greatly reduce the oppositions possibility of scoring. Plot B on the other hand is an example of an attacking action that results in a significant increase in the team's probability of scoring, but a smaller change in the probability of them conceding. Such an example would be if a player dribbled past the last defender and was 1-on-1 with the goal-keeper; the dribble would carry a similar *OBV* to this. Thus, though both actions achieve the same *Total OBV* of 0.17, they are performed in completely different situations in the game where one is critical for defensive reasons and the other for attacking reasons, but *OBV* allows them to be both valued equally.

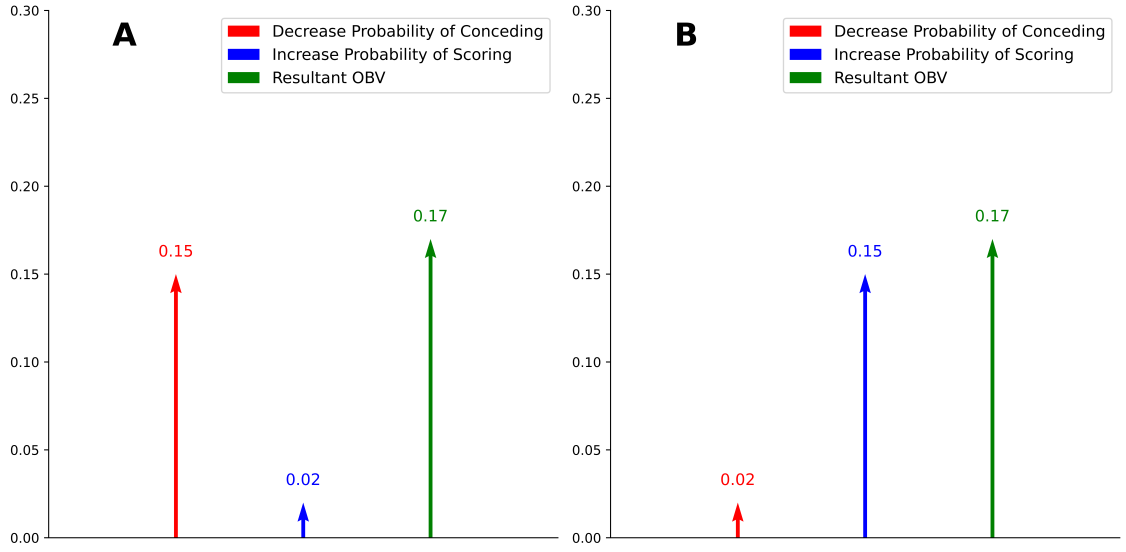

Figure 3: (A) Shows effect of a defensive action on *OBV*, while (B) shows effect of an attacking action on *Total OBV*

OBV can be broken up into its components, allowing for the types of actions valued to be understood. The equation for *OBV* is shown in equation (3.3.2.1) below.

Here we see Passing, Defensive Actions (DA) and Dribbles and Carries (DC) being awarded *OBV* scores. Large Pass- and *DC OBV* values are associated with passes and dribbles into good attacking areas, while significant *DA OBV* values deal with defensive actions—such as blocks, interceptions, clearances, or tackles—that either regain possession in a threatening attacking position or prevent the opponent from creating a significant scoring opportunity.

$$\text{OBV} = \text{Pass}_{\text{OBV}} + \text{DC}_{\text{OBV}} + \text{DA}_{\text{OBV}} \tag{3.3.2.1}$$

Notably, StatsBomb elected to employ two distinct models for delineating Goals for and Goals against components for each action, a methodological departure from both VAEP and xT. This choice separates the offensive and defensive impacts of individual actions. Another key feature is the ability of the model to discredit bad passes and thus negatively affect the players *Pass OBV* score upon an unsuccessful pass, something Statsperform's PV metric attempts, but xT and VAEP do not, as they only consider successful actions.

StatsBomb's *OBV* model is also trained on xG rather than goals scored. This approach acknowledges that xG is a superior predictor of future goals compared to actual goals scored, as well as using more data by considering all shots taken as opposed to solely the shots that lead to goals (octosport.io, 2022).

The *OBV* framework does not incorporate possession history features, which describe preceding events within a possession chain. This can be done to infer likely opposition defensive structures, as is adopted by the VAEP model. Although ideally this would allow for valuing actions while accounting for the game-state, incorporating these variables leads to autocorrelation with factors such as team play style and team strength. In turn, this leads to models overestimating the significance of passes occurring within longer possession chains, as stronger teams tend to sustain possession for extended durations compared to weaker counterparts (StatsBomb, 2021). Thus, *OBV* values each action independent of team strength, avoiding biases associated with possession history.

3.4 Decision Support Systems (DSS)

DSS in sports have gained significant traction due to their ability to aid and speed-up decision-making processes (Felfernig et al., 2024). These systems have a wide reach and assist individuals and organizations across numerous sporting domains by recommending suitable training programs, nutrition plans, strategies, and even talent. Some systems, like those implemented by Santos-Gago et al. (2019), focused on personalizing physical training sessions based on individual conditions and environmental factors, such as weather.

Additionally, DSS extend beyond traditional athletic contexts. In eSports, systems have recommended appropriate gaming maps, opponents, or even equipment tailored to player profiles, enhancing engagement and ensuring competitive performance (Wu et al., 2017). Nutrition-focused systems have also complemented these by suggesting performance-enhancing diets aligned with athletes' needs (Alcaraz-Herrera et al., 2022). Injury prevention systems have likewise played a critical role, having advised athletes on how to modify training loads to reduce the risk of injury (Peterson & Evans, 2019).

When it comes to recruitment and tactical planning, DSS are still evolving. In the recruitment space, knowledge-based recommenders systems have matched athlete profiles with specific team needs. For example, Rajesh et al. (2022) explored a system that assists in configuring football teams within the Fantasy Premier League by balancing constraints such as player budgets and position requirements. Furthermore, collaborative filtering methods were employed to great use by recommending substitutes by analyzing historical player data, using metrics such as passing accuracy and sprint speed (Yılmaz & Öğüdücü, 2022). This approach ensured that substitutes aligned with team strategy and maintained performance levels.

In football, DSS are applied in tactical and strategic contexts, though their use for player recruitment remains underdeveloped. Collaborative filtering have been used to help coaching staff identify successful tactics employed by other teams, offering insights into whether defensive or offensive strategies might be suitable for upcoming matches (Abreu et al., 2014). Similarly, content-based systems have been demonstrated to analyze video footage of past matches to provide recommendations on tactical adjustments (Wu et al., 2022). These insights were crucial in helping teams tailor strategies against specific opponents. Group DSS have additionally been employed to support tactical decision-making, which enabled coaches and staff to collaboratively decide on the best course of action for a match (Beal et al., 2019). An additional example of this type of in-game approach was outlined by El-Roby et al. (2023), where football analytics were integrated with historical data to improve real-time strategic decisions. Using Dynamic Time Warping (DTW) to compare ball movement patterns, the paper proposed a solution which identified similar ball movement patterns from large datasets, helping teams uncover tactical insights from past games. To manage the scale of historical data, they employed clustering to optimize computational efficiency. They also aligned relevant video snippets with event data through computer vision techniques, allowing coaches access to visual insights. This system not only improved the precision of in-game decisions but also ensured scalability and speed for real-time use.

While existing DSS in football span various domains - from nutrition and player recovery to pre-match and in-game tactical optimization - there remains a notable

gap in research focused on data-driven recruitment systems. This gap is particularly significant given the potential for sophisticated player evaluation metrics to help PSL clubs identify undervalued talent efficiently. This research therefore addresses an important intersection between advancing analytical methodologies and practical application in less affluent professional football teams.

3.5 Conclusion

The financial disparity between well-funded and resource-limited football clubs, both in South Africa and globally, significantly impacts the ability of financially constrained teams to remain competitive. As football continues to attract greater investment, this gap is likely to widen, reinforcing the advantage of wealthier clubs. Concurrently, the increasing reliance on data-driven decision-making has transformed recruitment strategies, with many clubs turning to data providers and analysts for competitive insights (Lolli et al., 2024). The iterative evolution of player evaluation methodologies, exemplified by clubs like Brentford and Brighton, further demonstrates the effectiveness of analytical approaches in modern football.

However, the cost of premium data acquisitions threatens to leave smaller clubs at a disadvantage, limiting their ability to compete at higher levels, avoid relegation, and qualify for prestigious competitions. Addressing this challenge requires the development of cost-effective scouting methodologies that can be implemented across clubs with varying budgets, ensuring that financially constrained teams are not left behind. In this context, OBV emerges as a particularly promising player evaluation metric, addressing key limitations of previous approaches. Given its comprehensive capture of on-ball contributions, OBV serves as a logical foundation for modeling, which in turn facilitates the development of a Decision Support System (DSS) designed to enable more accessible and scalable data-driven recruitment.

4 Exploratory Data Analysis (EDA)

The extent to which *OBV* could be effectively modeled using more affordable frequency statistics was unknown and was expected to be challenging, given the complexity of *OBV* and the simplicity of frequency-based metrics—particularly their lack of locational data. Another key consideration was whether to model *OBV* as a whole or to focus on its individual components.

Thus, the primary research questions explored in this section are: "Is *OBV* valid? That is, are teams with higher *OBV* generally ranked higher?", "Is there a significant relationship between cheaper datasets (Wyscout and FBref) and *OBV*?", and "Are *OBV* component models valid, or is *OBV* better modeled in its entirety?" Additionally, this analysis informed the basic model structure for the subsequent sections.

This section involved the implementation of three primary techniques. Principal Component Analysis (PCA) (Wold et al., 1987) was applied to reduce data dimensionality while preserving key patterns by transforming variables into uncorrelated components. t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) was also implemented to visualize high-dimensional data by creating a low-dimensional map that preserved relationships between points. Finally, Random Forest variable importance analysis (Breiman, 2001a) was applied to identify the most influential predictors of *OBV* by measuring how much each variable improved prediction accuracy. These methods are thoroughly detailed in the Methods section. Correlation analyses were also conducted to assess the feasibility of using linear models for *OBV* modeling, as their implementation would significantly reduce the complexity of establishing a scouting system for PSL clubs.

4.1 OBV Interrogation and Validation

4.1.1 What is OBV?

OBV is an all-inclusive metric designed to capture a player's quality on the ball, attempting to capture the change in the net likelihood of scoring due to each action. Figure 4 below showcases the distribution of average player *OBV* per match across the 3 seasons in the PSL and single Premier League season. Of particular interest was the normality of the distribution, as this informed whether we reliably use linear models to model *OBV*. This distribution however appears to be slightly skewed to the left, and a resultant p-value from a Kolmogorov-Smirnov test of normality of 0.0001 provides sufficient evidence that the distribution is not normal. Additionally, a kurtosis value of 1.324 indicated that the distribution is platykurtic, with a flatter peak and lighter tails compared to a normal distribution, further highlighting

deviations from normality.

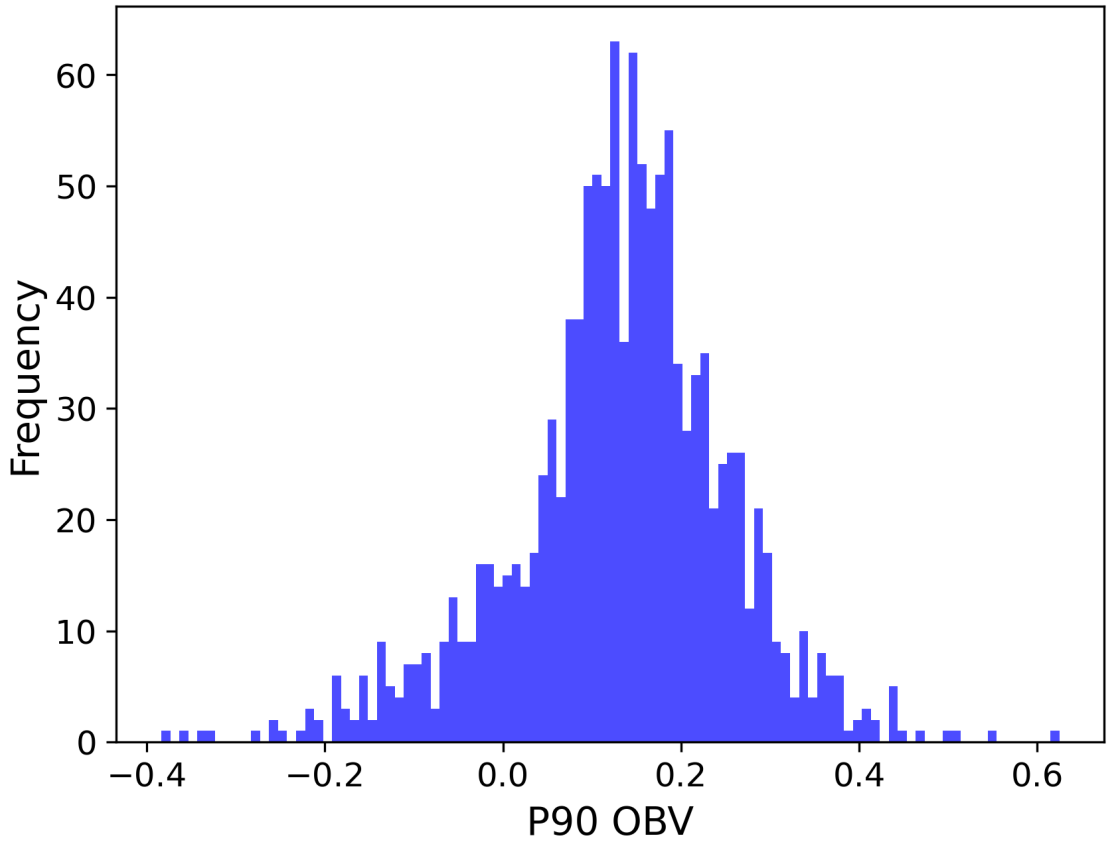

Figure 4: OBV over 4 seasons of Statsbomb data

Table 1 below further describes the distribution shown above. The OBV distribution had a median value of 0.14. The range was very large, with a value of 0.62 for the maximum OBV achieved by a player, and -0.38 for the minimum. This showed the large disparity between top-performing players and bottom-performing ones existed. Interestingly, of the top 5 performing players, 3 were fullbacks, hinting that this could be a high value position.

Statistic	Value
Minimum	-0.3837
First Quartile	0.0720
Median	0.1403
Third Quartile	0.2038
Maximum	0.6250

Table 1: Summary statistics of P90 *OBV*

Since *OBV* measures the change in the net likelihood of scoring, presumably attacking actions (passes and carries) which move the ball towards the opposition goal are likely to generate more *OBV* than similar actions which move it away. Additionally, it was expected that defensive actions closer to one's own goal are likely to garner more *OBV* as they mitigate a greater threat than defensive actions in the middle of the pitch. Figure 5 below illustrates a few passes which show a range of *OBV* values.

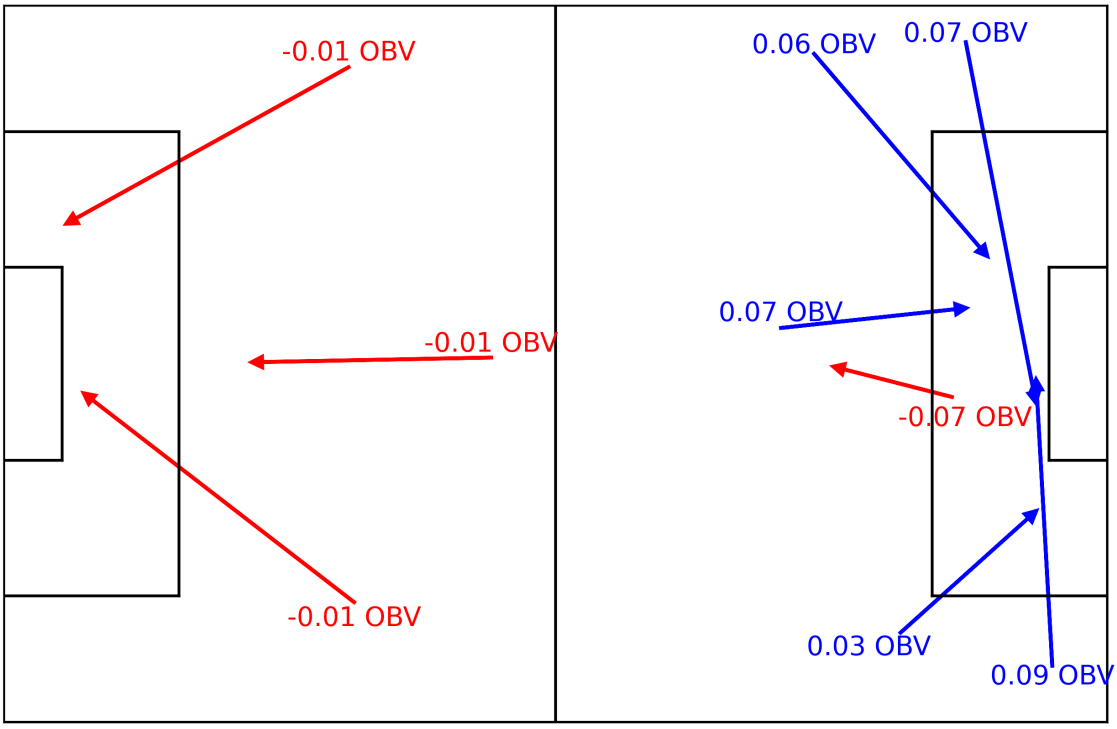

Figure 5: Successful Passes with their respective *OBV* values

Starting on the left, the 3 passes in the defensive half are all examples of passes back

to the goalkeeper, typical passes in football. Even though these are over similar distances when compared to passes in the other half, they achieve small absolute values. This is because they are not in areas likely to lead to a goal. Passes into the box however are seen to be very highly rewarded, with values such as 0.09, and 0.07 awarded to the crosses into the middle of the box. This area appears to be from where it is most likely to score as the two passes into the box are given the largest OBV values. Similarly, passes from inside the box (high scoring probability zone) to outside the box (lower scoring probability zone) are penalised, highlighting the effect of moving the ball away from a high goal scoring region.

Figure 6 below shows similar results for dribbles as well. Once again, we see negative values attributed to dribbles towards one's own goal, while dribbles into the box are given strong positive values, as these are seen to strongly increase the probability of scoring. Dribbles which take the ball from within the box to outside the box are once again negatively scored, as they decrease the probability of scoring, as one has gone from an area of high threat to an area of lower threat.

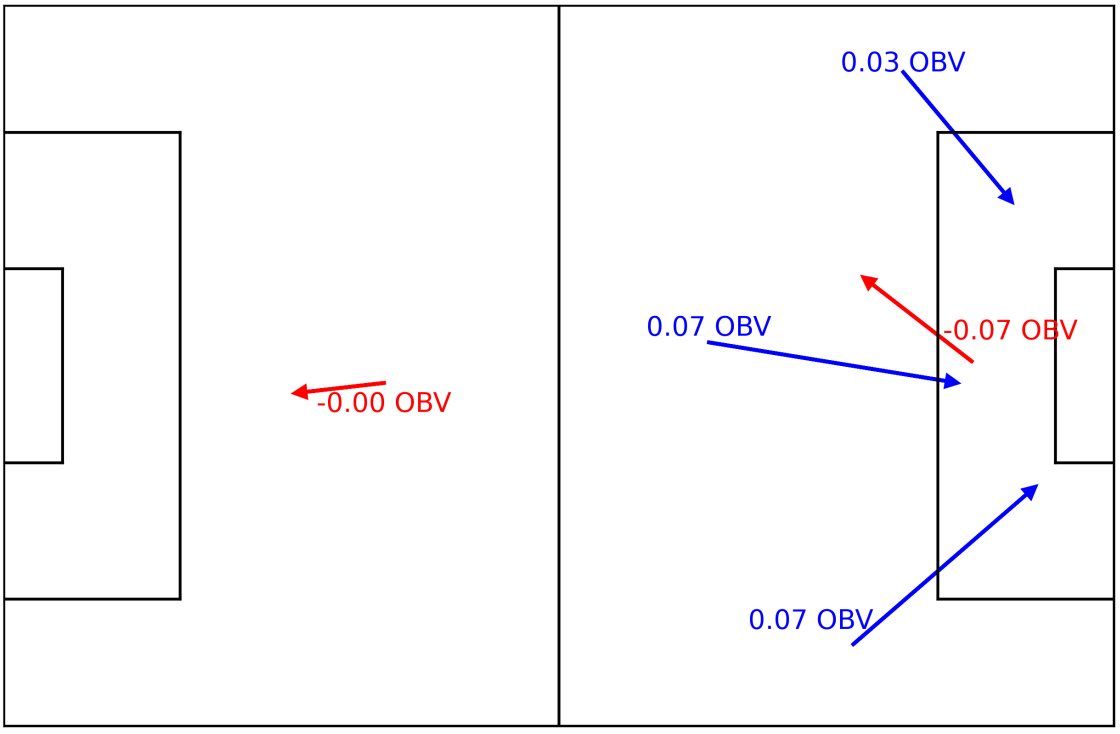

Figure 6: Successful Dribbles with their respective OBV values

The final component of OBV is defensive actions. This relationship is easily seen

through the (Figure 7) below. The heatmap is weighted by average *DA OBV* per grid cell and shows that defensive actions inside the box—whether in a player's own box (left) or the opponent's box (right)—yield the highest *OBV*, with values diminishing further from the box. Hence, it was noted that to predict *DA OBV* accurately, clearly the location of the defensive actions was needed, as the value of the *DA OBV* is clearly linked to where the action takes place. Furthermore, the level of locational granularity required to adequately model this output variable was evident, as even within the same box the level of *DA OBV* differs largely between right in front of the goals and the corners of the box.

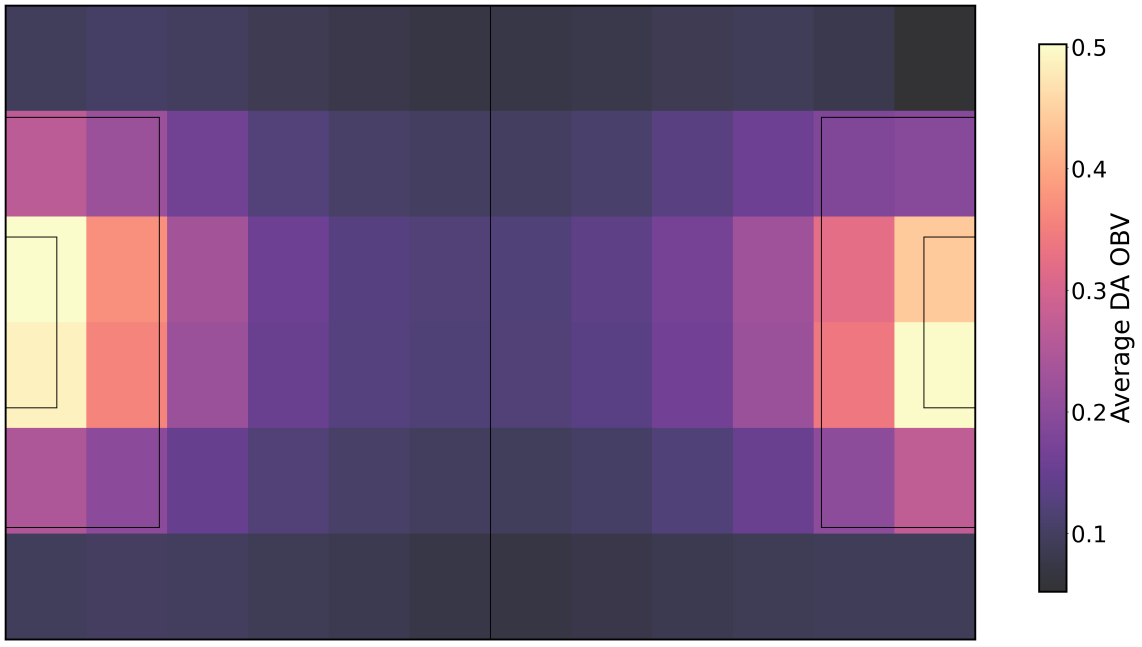

Figure 7: Heatmap of average *OBV* from defensive action in each cell

4.1.2 Validation OBV

One of the main strengths of *OBV* is its ability to capture a holistic measure of a player's quality, furthering a team's ability to value players and thus assess how good a team is quantitatively. Thus, *OBV* is a strong candidate metric to use when scouting players. The idea of scouting is to purchase players who will improve the team's performance and consequently improve the team's position on the log. If a team could buy points, that would be perfect because points are trivially perfectly correlated with a team's position finished.

Goal difference is another variable highly correlated with log Position (Figure 8 below). This suggests a strategy of buying players who have the greatest impact

on a team's scoring potential and on preventing the team from conceding goals. However, when buying players who contribute to scoring, this would likely result in mostly strikers being bought, since their primary role is to score. In addition. how can one quantify a player's contribution to a goal conceded? This is where *OBV* becomes helpful.

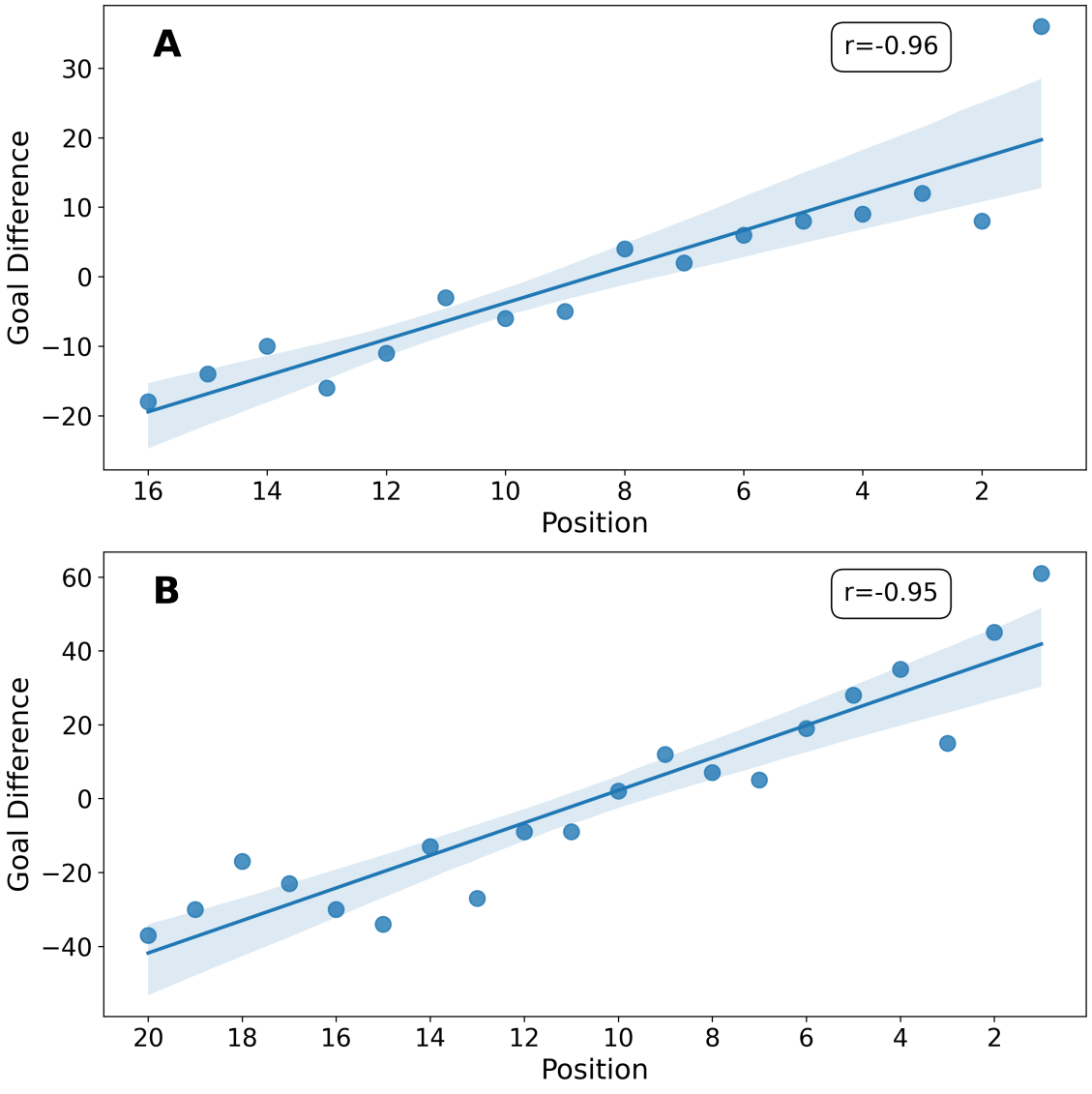

Figure 8: Correlation between goal difference and final league position where subplot A is the PSL (21/22) and B is the English Premier League (22/23)

As discussed previously, *OBV* captures a player's contribution to a team scoring and

conceding, allowing for both attacking prowess and defensive output to be valued, however the strength of the relation between a team's OBV and their league finish had not yet been quantified. Figure 9 below addresses this issue. A strong positive relationship was observed between a team's OBV rank and their position on the table, with Spearman correlation coefficients of 0.72 for the PSL 21/22 season and 0.75 for the Premier League. This indicated a clear association between higher OBV levels and better table rankings. Even though this association is weaker than that of goal difference with league position, OBV can be indirectly bought through player acquisitions, while goal difference cannot. This implied that OBV should be an instrumental metric when scouting players, allowing for an efficient, objective valuation of players.

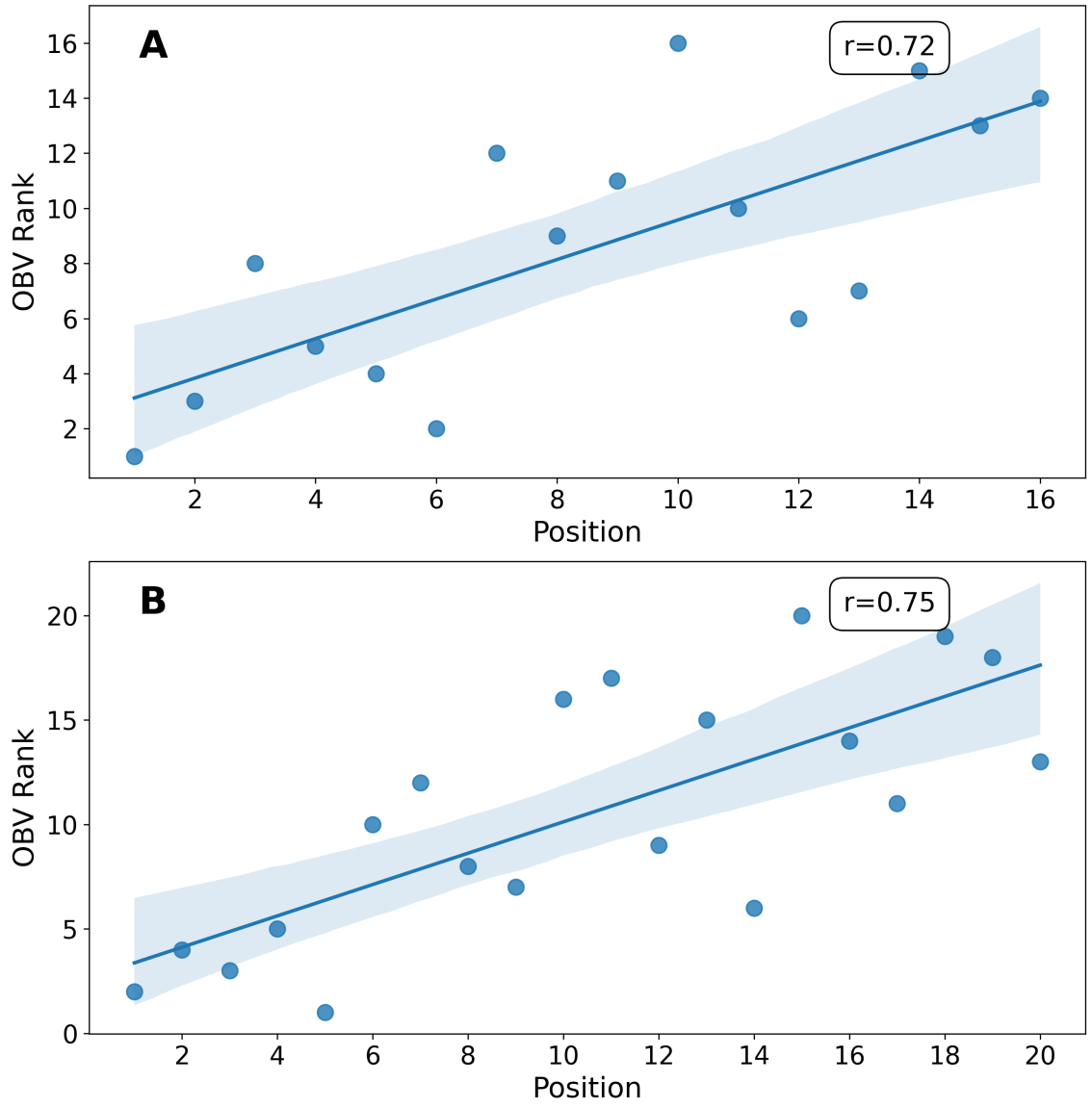

Figure 9: Correlation between final league position and *OBV* rank where subplot A is the PSL (21/22) and B is the English Premier League (22/23)

4.1.3 OBV Exploration

Analysing cumulative *OBV* by position can reveal where teams are on average extracting most of their value, with better teams presumably extracting more than average from these positions. In this analysis, the following abbreviations for player positions are used: Center backs (CB), Fullbacks (FB), Midfielders (M), Wingers

(W), and Forwards (F). Figure 10 below shows the *Total OBV* per position category accumulated over the 4 seasons of data, and indicated the significance of midfielders and fullbacks.

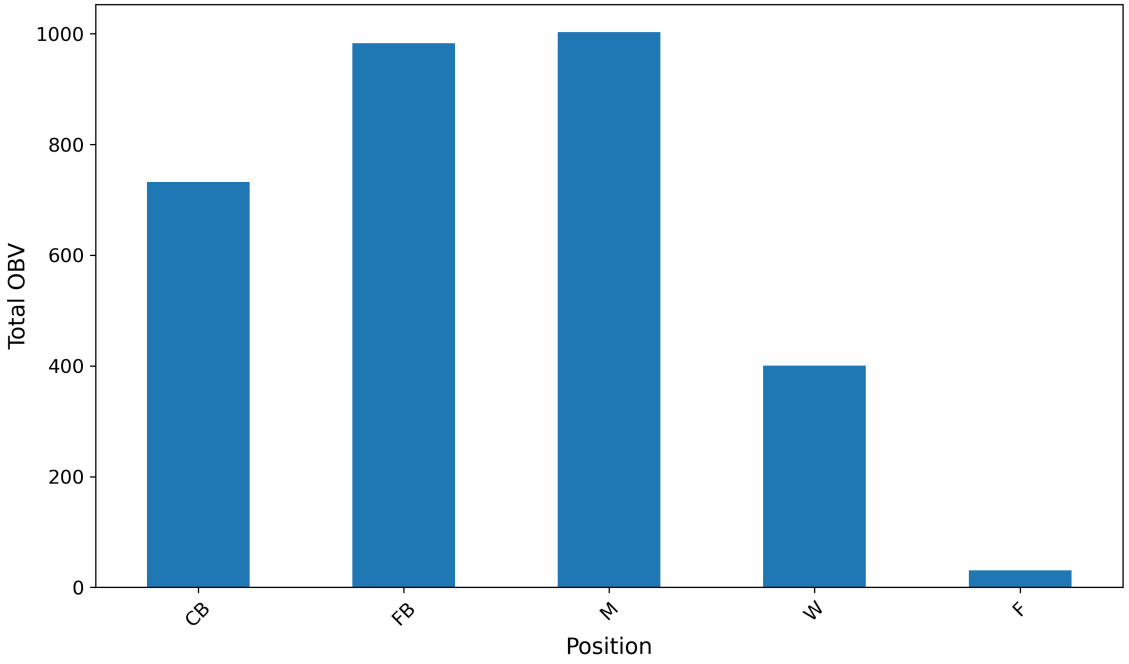

Figure 10: Cumulative *OBV* by position

In Figure 10 Midfielders are seen to create the most *OBV*, while fullbacks trail slightly behind them. Fullbacks and Midfielders are found to have the most significant *OBV*, which was somewhat expected as midfielders generally outnumber the other positions and fullbacks are often advanced up the pitch in the modern game (Konefal et al., 2015). Interestingly though, wingers are found to have less *OBV* than fullbacks, even though fullbacks are generally found behind wingers on the pitch, and thus are generally less likely to generate *OBV* from their position due to it being a less dangerous position on the pitch. Forwards garner very little *OBV* which is to be expected they also make a great deal of back passes (see Figure A1 in the appendix which illustrates that the further up the pitch you play the more back passes you tend to make, both of which will decrease an individual's *passing OBV*). This Figure is susceptible to biases and doesn't give the full picture of *OBV* however, since different numbers of players operate in each position. To see which position on average generates the most *OBV*, we needed to analyse the average *OBV* generated per game, or roughly per 90 minutes, (P90 *OBV*) per position.

Analyzing *Total OBV* per 90(A), fullbacks lead followed by wingers and midfielders,

while forwards show minimal values, reflecting fullbacks' dual offensive-defensive role. Looking at *DC OBV* per 90(B), wingers dominate in dribbles and carries, with midfielders and fullbacks also showing significant involvement in ball progression. Center backs lead in *DA OBV* per 90(C), followed by fullbacks and midfielders, which aligns with their crucial defensive responsibilities. In *Pass OBV* per 90(D), fullbacks and center backs show the highest values, highlighting their important role in ball distribution and build-up play.

The mean values (displayed in Figure 10 below) were useful for understanding the average impact of a player in each position. However, violin plots provided a richer view of how the data is distributed than standard bar plots. These plots combine box plot elements with a kernel density estimation that creates a mirrored density shape, resembling a violin. The width at any point shows how common that value is in the data, while the overall shape reveals the full distribution of *OBV* components across positions. This allowed us to see not just averages, but also patterns like whether the data is skewed, has multiple peaks, or contains outliers. Through violin plots, we could better understand how *OBV* components varied by position. Figure 11 below shows this.

Subplot A shows that fullbacks have the largest *Mean OBV* per 90 minutes, followed by wingers and midfielders. Center backs display a lower and more concentrated distribution, with forwards showing a very limited range. This indicated that fullbacks are likely significantly involved in generating offensive and defensive value, reflecting their dynamic role, as well as highlighting a limitation of *OBV* in that the variance in forwards' *OBV* is minimal.

Subplot B highlights that wingers exhibit the highest and largest variance of *DC OBV* per 90, with midfielders and fullbacks also showing substantial contributions. This underscored the pivotal role wingers play in advancing the ball through dribbles, breaking down defensive lines and creating opportunities. The lower mean values for center backs and forwards suggests less valued dribbles occur through players in these positions, since the width of the violin plot is larger than that of wingers.

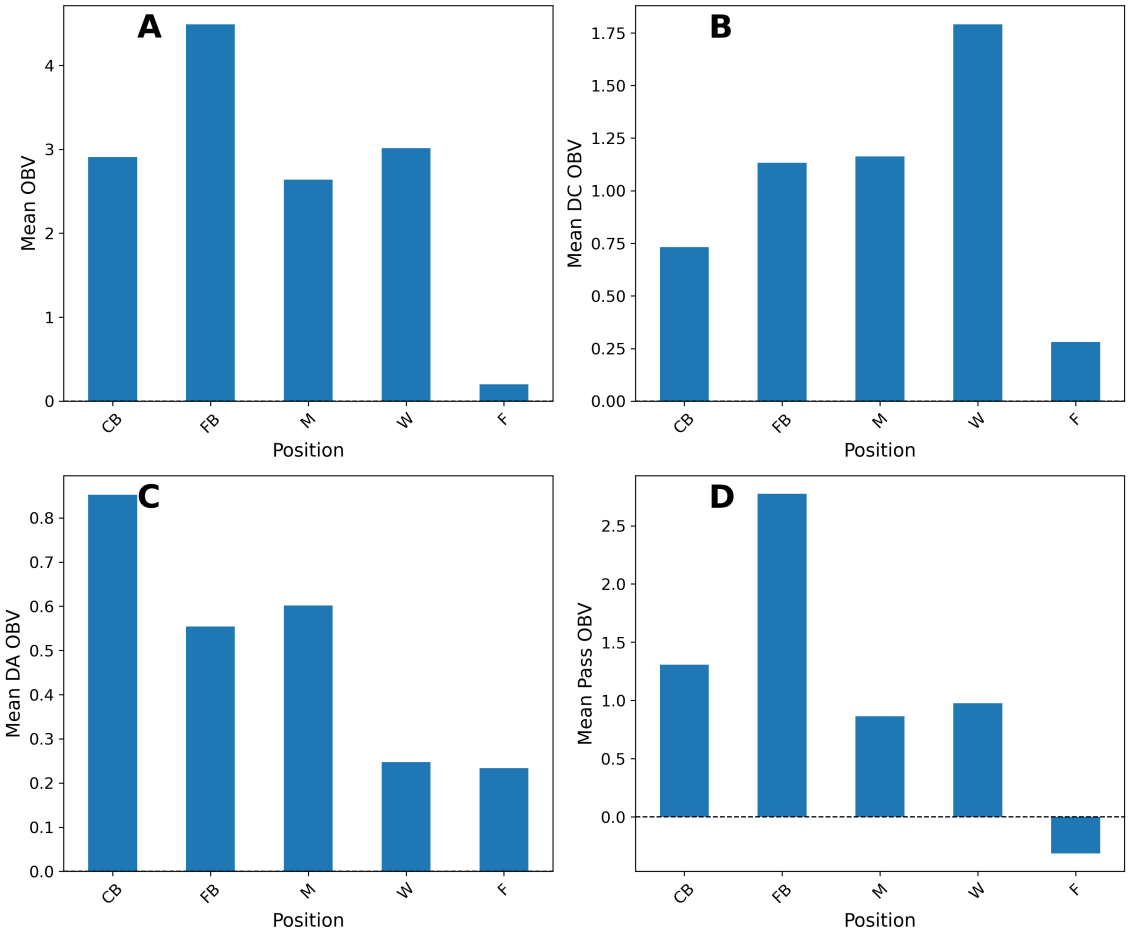

Figure 11: *OBV* split by components and positions

Subplot C shows that center backs have the highest and most varied distributions in *DA OBV* per 90, followed by fullbacks. This was expected as center backs are generally the last line of defence and the heatmap (Figure 7) showed the highest *DA OBV* actions occur closest to the goal. Center backs who frequently miss tackles in dangerous areas near their goal show significant negative *DA OBV* values, as these defensive errors increase the likelihood of opponents scoring.

The distribution of *Pass OBV* by Position Category Per 90 is seen in subplot D, and full backs are seen to lead with the highest average value, as well as a very wide distribution, showing some fullbacks contribute significantly with their passing, and others do so far less. This is to be expected, as we saw crossing balls into the box is an action that can achieve a high *Pass OBV* score, and so full backs who don't cross (more defensive minded fullbacks) will be significantly less valued than those who

do. The distributions for midfielders, center backs, and forwards are centered at a lower average value, which suggested a generally less valuable role in high-performing passing sequences, when compared to full backs.

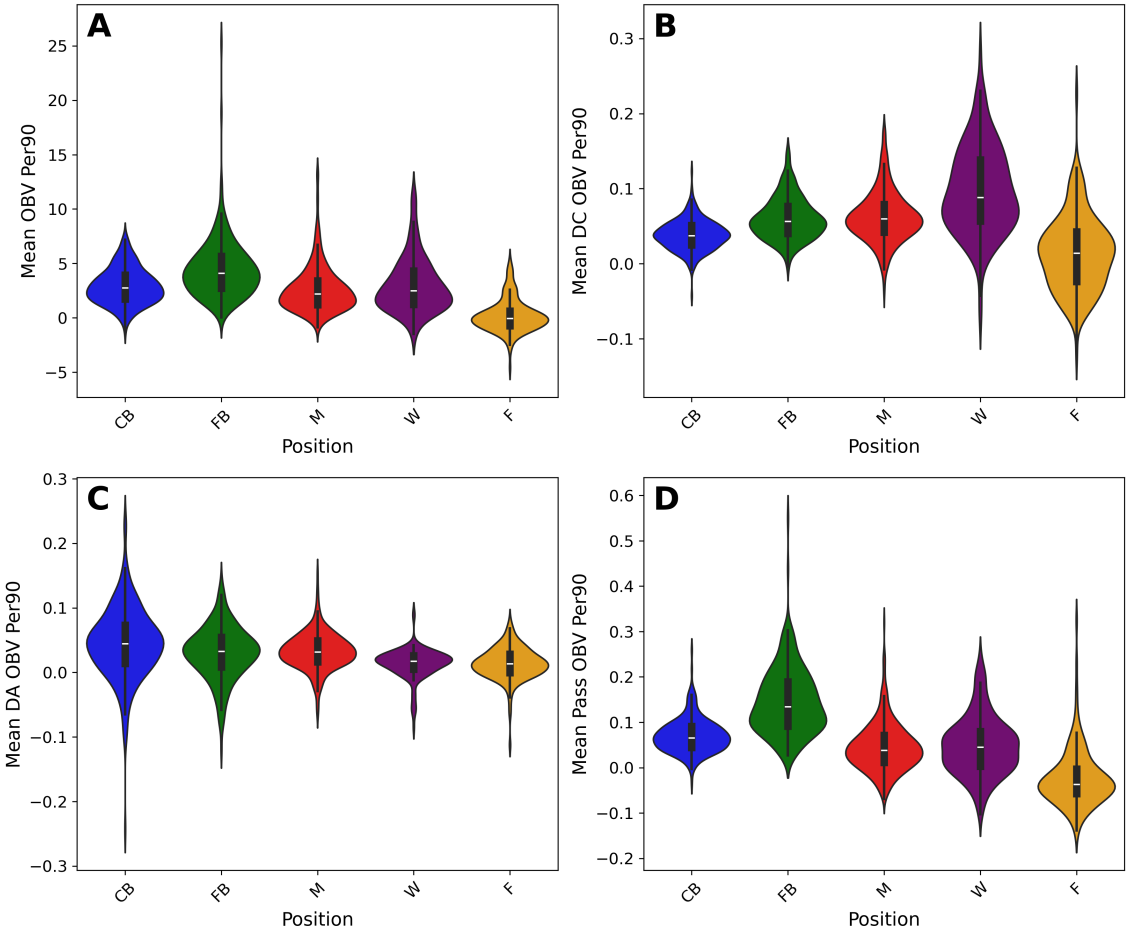

Figure 12: Distribution of *OBV* components per position

Figure 13 below shows the distributional differences of the three *OBV* components: *Pass OBV*, *DC OBV*, and *DA OBV*. All three *OBV* components are narrowly distributed above zero, with *Pass*- and *DC OBV* having mean values of 0.061 and 0.054 respectively, while *DA OBV* is centered at a lower *OBV* value of 0.03. The variance across the distributions was particularly intriguing, with *Pass OBV* showing the highest variance at 0.0058, while *DC OBV* and *DA OBV* exhibit much lower variances at 0.0019 and 0.0015, respectively. This is underscored by the inter-quartile range being roughly double for the *Pass OBV* distribution, compared to that of the *DC* and *DA*- *OBV* distributions. This finding showed the importance of evaluating

OBV component prediction models using more than just error metrics like MSE and MAE. Given the low variance in the *DC*- and *DA OBV* components, models will be able to perform better on these components compared to the passing model with respect to error metrics, without necessarily having a greater understanding of the underlying relationships. This is why it was crucial to also consider metrics which quantified the model's explainability of the output variable (e.g. adjusted R^2) as they provide deeper insight into how well the model captures the relationships between the output and the input variables.

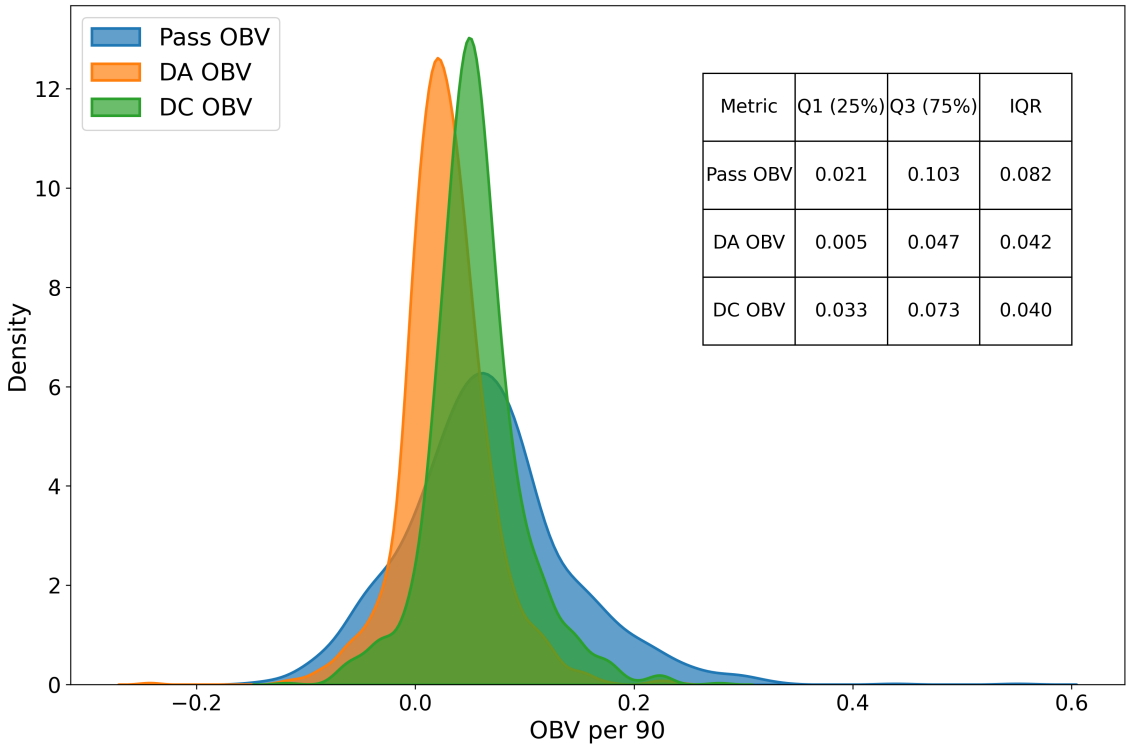

Figure 13: Density plots of *OBV* components

4.2 Modeling OBV

Modeling *OBV* involved a detailed examination of various explanatory variables to determine their effectiveness in explaining *OBV*. There were two primary sources these explanatory variables were taken from: Wyscout (Wyscout, 2024) and FBref (FBref, 2025). The Wyscout dataset encompassed all the explanatory variables needed to model the 4 seasons of *OBV* data, namely the 3 PSL seasons and the single Premier League season, however the FBref dataset only had data for the Top 5 European leagues, so only the Premier League *OBV* data (1 season) could

be considered for this dissertation. The fundamental challenge was to utilise non-locational-based variables (the explanatory variables) to model something that was fundamentally based on location (OBV).

This analysis had three main objectives: identify the key factors driving OBV, determine if non-spatial variables can adequately account for OBV's spatial aspects, and evaluate whether linear models are appropriate. To achieve this, relevant variables were first identified and selected from the Wyscout and FBref datasets through an intercorrelation analysis. Additionally, a correlation analysis was performed between the input variables and OBV to better understand the validity of linear prediction models. Techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE) were then employed to visualize and understand the relationships between potential predictor variables and their contributions to OBV. This also provided insight into the potential utilization of linear and non-linear dimension reduction techniques. Random Forest variable importance was subsequently used to identify the most significant variables in explaining OBV, offering insight into which variables held the greatest predictive power and enabling potential dimensionality reduction. Finally, position-based interaction terms were analyzed to assess how they could better capture spatial relationships in the data, allowing for an evaluation of the suitability of linear models for this analysis.

4.2.1 Correlation Analysis

This section delves into an analysis of the variables available in the Wyscout and FBref datasets, with a particular emphasis on identifying correlations among these variables. Correlation in variables can lead to issues such as multicollinearity, which can distort a model's ability to correctly identify the relationships between the predictors and the response variable, in this case, OBV, lowering the certainty of any possible predicted relationships.

4.2.1.1 Wyscout dataset

The correlation analysis revealed significant redundancy among Wyscout variables, particularly in passing and defensive metrics. Several passing variables showed high correlations (≥ 0.7), which lead to the removal of basic metrics (*Passes.per.90*, *Assists.per.90*) in favour of more contextually relevant ones like *Progressive.passes.per.90* and *xA.per.90* (expected assists). In defensive metrics, correlations exceeded 0.9 between raw and possession-adjusted versions of the same actions, such as that between *Interceptions.per.90* and *PAdj.interceptions.per.90*, resulting in the retention of only possession-adjusted metrics, as these adjust for the frequency bias inherent in defensive actions, since players whose teams have possession of the ball less are more likely to make interceptions, since their team are trying to get the ball back. Dribbling

variables showed minimal redundancy and were all retained due to their distinct contributions. A detailed and comprehensive correlation analysis can be found in the Appendix in Section A.4.1.

4.2.1.2 FBref dataset

The FBref dataset also contained significant overlap amongst the features, leading to numerous variables being removed. In the passing category, several variables showed high correlations, particularly among basic metrics (*Medium.passes.per.90*, *Shot.creating.actions.per.90*), which again lead to their removal in favour of more location-based variables such as *Passes.into.final.3rd.per.90*. Additionally, new engineered features such as *non.progressive.passing.distance.per.90* were created to capture unique aspects of passing performance. The defensive variables displayed only weak to moderate correlations and were therefore all retained. Similarly, while dribbling variables showed some correlation, particularly between *Carries.per.90* and *Progressive.carry.distance.per.90*, all were kept as they measured distinct aspects of ball progression. The full correlation analysis can be found in the Appendix in section A.4.2.

4.3 Comparing Total OBV with Wyscout and FBref variables

Figure 14 below displays the number of players in each position within the Wyscout and FBref datasets.

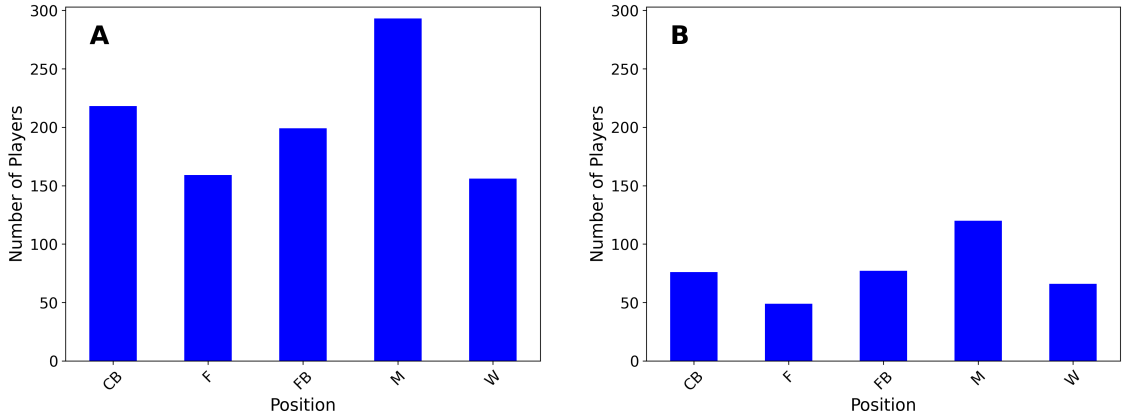

Figure 14: Comparison of player counts across positions in Wyscout (A) and FBref (B) datasets

The Wyscout and FBref datasets showed similar positional patterns but differed

significantly in scale, with the Wyscout dataset containing 1025 observations and the FBref dataset only 388. Consequently, the Wyscout dataset had larger sample sizes across all positions, with midfielders dominating both datasets (293 in Wyscout, 120 in FBref). This high midfielder count aligned with what Beernaerts et al. (2022) found regarding formation patterns. The relative proportions remained consistent between datasets. The FBref dataset, which consisted of just one season of the Premier League, had the lowest representation of forwards (49).

Table 2 below presents the five variables most highly correlated with OBV per 90 minutes in both datasets. This allowed us to observe how strongly each variable was linearly associated with a player's OBV per 90, suggesting how important each variable could be in linearly modeling OBV in the later sections.

Wyscout Dataset		**FBref Dataset**	
Variable	**Correlation**	**Variable**	**Correlation**
Progressive. passes.per.90	0.635	Crosses.into penalty.area	0.510
Aerial.duels. per.90	0.468	Long.passes per.90	0.510
Average.long.pass. length.m	0.441	Passes.into. penalty.area.per.90	0.495
Passes.to.penalty. area.per.90	0.428	Expected.assists. per.90	0.490
Duels.won. percent	0.415	Progressive. carries. distance.per.90	0.418

Table 2: Top variable correlations with OBV in the Wyscout and FBref datasets

In both datasets, passing-related metrics showed the strongest correlations with OBV. The Wyscout dataset revealed *progressive.passes.per.90* as the highest correlated variable (0.635), while in FBref, *crosses.into.the.penalty.area* (0.510) and *long.passes.per.90* (0.510) had the highest correlations. Both datasets demonstrated the importance of passes into the penalty area, with correlations of 0.428 (Wyscout) and 0.495 (FBref), suggesting a consistent valuation of passes into dangerous areas across data sources.

The datasets differed in their secondary indicators: Wyscout exhibited strong correlations with aerial metrics (*aerial.duels.per.90* at 0.469) and *Average.long.pass.length.m* (0.441), while FBref emphasized *Expected.assists.per.90* (0.49) and *progressive.carries.-distance* (0.418) as significant contributors. Notably, FBref had higher correlations

with carrying metrics, while Wyscout showed a stronger relationship with the defensive action *Duels.won.percent* and *OBV*.

Both datasets suggested that linear models could be used to predict *Total OBV*, given the moderate correlations across multiple variables. However, they shared a common limitation: while passing metrics were well-represented and likely to be modeled effectively, both datasets showed limited correlations with defensive actions and dribbling components of *OBV*. This suggested that modeling these specific aspects of *OBV* might require more sophisticated, non-linear approaches. The findings across both datasets supported the implementation of simpler and therefore more easily interpretable linear models in scouting departments, particularly in South African clubs where resource constraints may exist. However, further analysis of variable relationships with specific *OBV* components was necessary to assess the potential for developing more comprehensive models.

The well-known dimensionality reduction method termed Principal Component Analysis, or PCA (Wold et al., 1987), was conducted on the datasets, producing the scree plots below. The scree plots from both Wyscout and FBref datasets revealed remarkably similar patterns in their principal component analysis. Both datasets exhibited a steep initial increase in cumulative explained variance for the first few principal components, with a characteristic elbow appearing at the 3rd PC in each case. However, this similarity extended to their limitations as well—at 3 PCs, neither dataset achieved 50% explained variance, indicating that simple linear dimensionality reduction may be insufficient when modeling *OBV*.

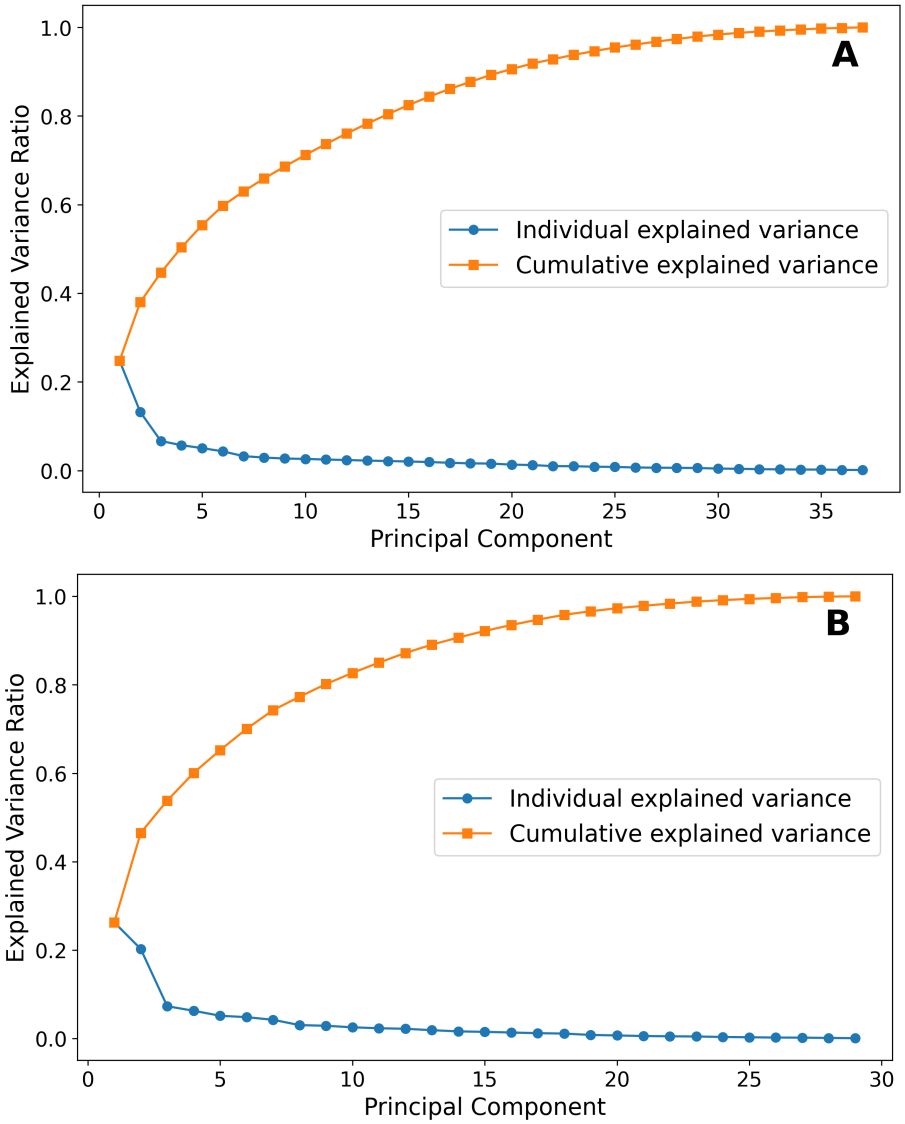

Figure 15: Screeplots of Wyscout (A) and FBref (B) datasets

Both datasets demonstrated that a substantial number of PCs (approximately 15–20) would have been necessary to capture 90% or more of the variance. This finding suggested that non-linear dimension reduction methods might be appropriate for effectively reducing the dimensionality of either dataset while preserving their essential information. Figure 16 below presents two PCA biplots, each overlaid with a Kernel Density Estimation heatmap to reveal underlying patterns in the data distribution. The heatmaps differ in scale between Plot A and B due to variations in data point density affecting the smoothing algorithm.

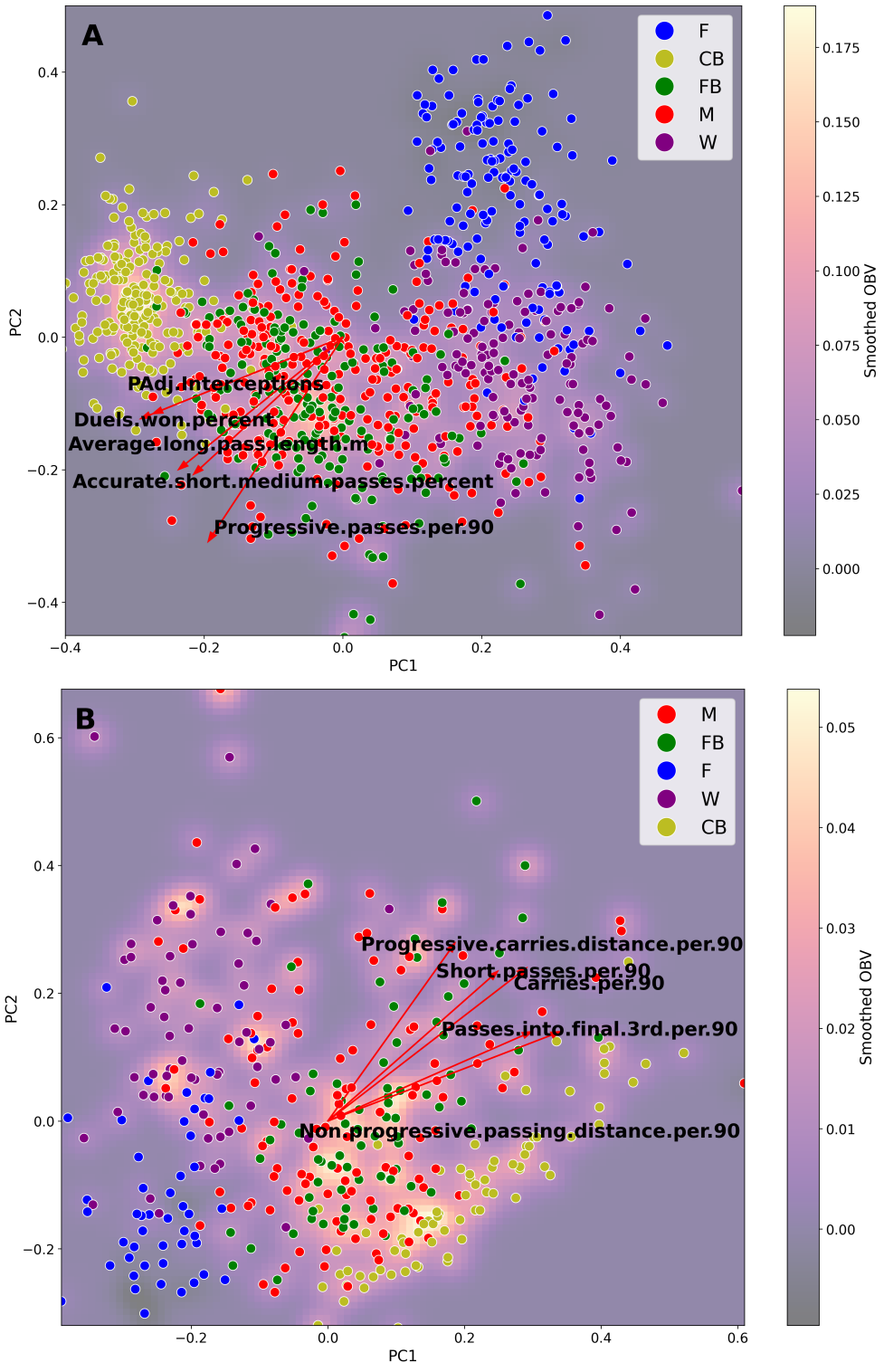

Figure 16: PCA biplots on all Wyscout (A) and FBref (B) variables overlaid onto an OBV heatmap with the five labelled variables which explain the greatest variance

Figure 16 above reveals that both Wyscout and FBref datasets faced similar challenges in their PCA results, particularly in distinguishing between player positions. In both cases, there was significant overlap between fullbacks and midfielders, while center backs emerged as the most distinctly separated position. The primary difference lay in the variables driving these separations.

The principal component loadings can be found in Table A7 and Table A8 in the Appendix. The first principal component (PC1) in both datasets appeared to represent a pitch-length dimension due to the dispersion of positions in Figure 16 A and B above; however, it was achieved through different metrics. While Wyscout's PC1 emphasized the contrast between defensive and creative actions, with negative loadings for center back-associated metrics (*Average.long.pass.length.m, duel.win.percent*) and positive loadings for attacking metrics (*progressive.runs.per.90, accelerations.per.90, smart.passes.per.90*), FBref's PC1 focused more on passing dynamics. This resulted in positive loadings for progressive passing metrics (*Progressive.passing.distance.per.90*) and negative loadings for *Take.ons.attempted.per.90*. The second principal component (PC2) also showed some distinctions, as Wyscout's PC2 differentiated players based on ground versus aerial involvement, with negative loadings for progressive and short-medium passes contrasting with positive loadings for aerial duels. FBref's second PC concentrated more on attacking dynamics, with positive loadings for carries and passes into critical areas, though its negative loadings (*Take.ons.succeeded.per.90, Tackles.in.defensive.3rd.per.90*) were less definitively interpreted.

Both datasets' PCAs demonstrated limited effectiveness in capturing OBV patterns, as evidenced by the kernel density heatmaps showing minimal alignment between variable loadings and high-performing players. This finding suggested the need for non-linear dimension reduction techniques to better capture the complexity of player roles and performance metrics.

The plots in Figure 17 below show four plots where the non-linear dimensionality reduction method t-distributed stochastic neighbor embedding (t-SNE) was employed. Plots A and C were colored by three broad position categories (Defender, Midfielder, and Forward), while Plots B and D were colored by five positions (Center Back, Fullback, Midfielder, Winger, and Forward). T-SNE was performed on the Wyscout dataset in Figures 17 A and B, and on the FBref dataset in Figures 17 C and D.

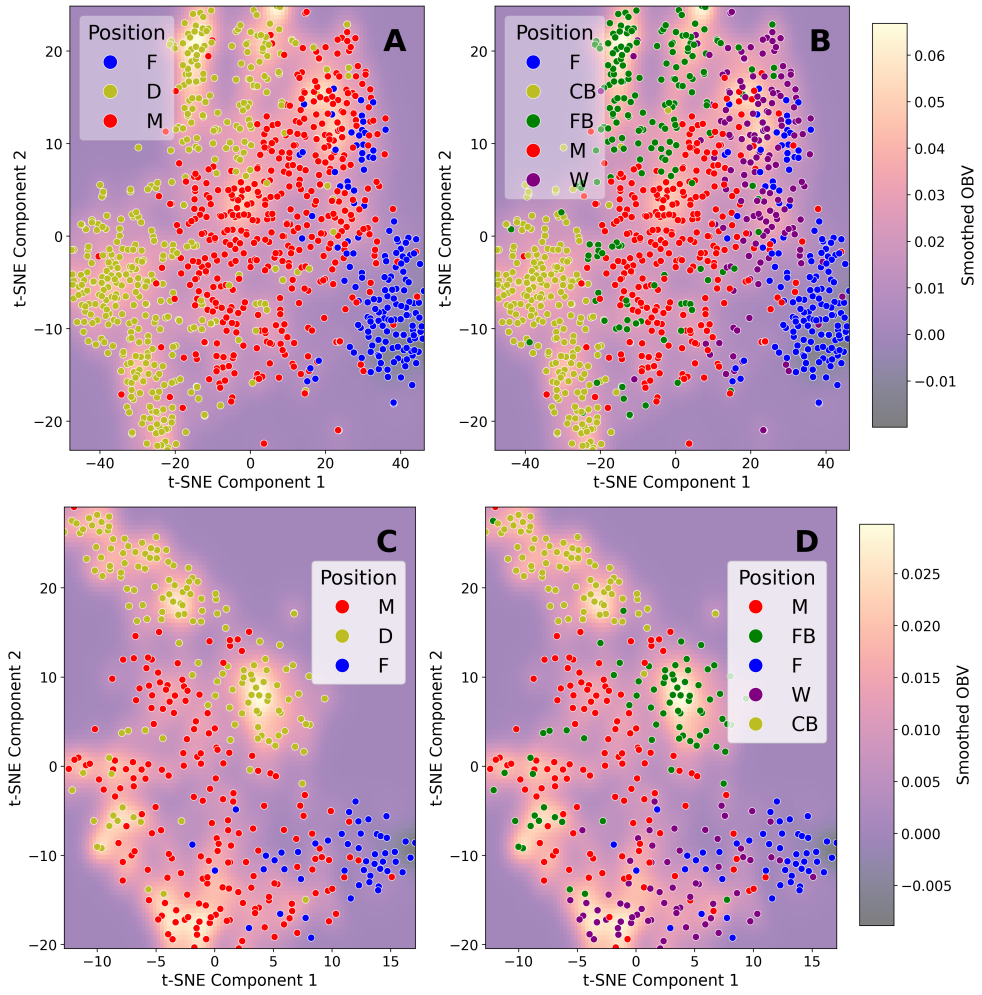

Figure 17: First two t-SNE components from Wyscout (A, B) and FBref (C,D) data overlaid onto an total OBV heatmap

Figure 17 shows that Wyscout and FBref datasets exhibited markedly improved clustering through t-SNE compared to their PCA results, though with some distinct spatial orientations. In Wyscout, the field length mapped horizontally (defenders left, forwards right), while FBref displayed it vertically (defenders top, attackers bottom). Thus, increasing values of t-SNE Component 1 coincided with players who were more attacking in the Wyscout data, while in the FBref data, decreasing values of t-SNE Component 2 tended to follow an increase in the attacking nature of the player. Despite this orientation difference, both analyses revealed similar patterns. Both datasets showed natural position-based clustering, with defenders forming distinct groups (three clusters in both datasets). Fullbacks emerged as a

unique group in both datasets. Their distinct clustering reflected that their unique combination of on-field actions and frequencies described a playing style different from defensive and midfield positions. Notably, both datasets showed fullbacks to be closer to the wingers and progressive midfielders than center backs were, thus highlighting their more attacking nature. Midfielders demonstrated similar fluid positioning in both analyses, scattered between defenders and attackers. This was likely because some midfielders were more defensive than others and had similar playing styles to defenders, while others were more attacking players and were thus found closer to the wingers and forwards. Wingers were generally found among the attacking midfielders and forwards, suggesting their role was primarily in progressive, attacking sequences of play.

With regards to *OBV* performance patterns, both analyses revealed that t-SNE clustered high-performing players together moderately well, with the FBref dataset producing more of these clusters, despite *OBV* not being used in the clustering process. Both datasets clustered the high-performing fullbacks, while FBref notably identified a specific high-performing cluster of mainly wingers at the bottom of the plot.

The ability of t-SNE to group similar playing styles and high-performing players strongly supported the use of non-linear methods for both dimension reduction and subsequent *OBV* modeling. The fact that these patterns emerged in both datasets, despite the FBref dataset's smaller size, suggested that non-linear approaches were more effective at reducing the inherent complexity of football performance metrics to a smaller set of features when compared to linear methods like PCA.

Having thoroughly analysed *OBV* across both the Wyscout and FBref datasets, the limitations of linear reduction methods and the potential strengths of non-linear dimension reduction approaches became evident. This analysis revealed that playstyle trends typically associated with specific positions were apparent in the data, indicating that incorporating a player's position could capture the locational aspects of *OBV* in the modeling process. Additionally, the relation between individual variables and *OBV* suggested the potential to model *OBV* using linear models, which would be advantageous due to their simplicity and ease of implementation for club analysts. With this understanding, the next step was to isolate the individual components of *OBV*.

By breaking down *OBV* into its components—passing, dribbling, and defensive actions—it became possible to examine each component separately to understand their relationship with the explanatory variables from the two datasets. This detailed analysis allowed for the identification of which components of *OBV* could be most robustly modelled by explanatory variables and further assessed the validity of using

linear methods to model these *OBV* components.

4.4 Comparing OBV Components with Wyscout and FBref variables

In this section we analysed the Wyscout and FBref variables and their relationship with each corresponding *OBV* component. The potential for *OBV* to be modelled using the full Wyscout and FBref dataset had already been demonstrated, but here we investigated the potential to model each component individually as that would allow for a more granular understanding of what components players are strongest in, thus allowing for a more in-depth scouting procedure.

4.4.1 Pass OBV

4.4.1.1 Pass OBV Correlation Analysis

Table 3 below shows the five passing variables with the highest correlation with *Pass OBV* for both the Wyscout and FBref Datasets.

Wyscout Dataset		**FBref Dataset**	
Variable	**Correlation**	**Variable**	**Correlation**
Progressive.passes .per.90	0.692	Long.passes .per.90	0.638
Average.long.pass .length.m	0.474	Crosses.into .penalty.area	0.558
Passes.to.penalty .area.per.90	0.456	Progressive.passing .distance.per.90	0.502
Average.pass .length.m	0.425	Expected.assists .per.90	0.462
Passes.to.final .third.per.90	0.411	Passes.into .penalty.area.per.90	0.432

Table 3: Top 5 variable correlations with *Pass OBV* in the Wyscout and FBref datasets

Both datasets show stronger correlations when focusing specifically on *Pass OBV* compared to their correlations with *Total OBV*, suggesting that isolating this component reduced noise from other *OBV* components. In Wyscout, *Progressive.passes.per.90* emerged as the strongest correlator (0.692), showing a slightly higher correlation

than it did with *Total OBV* (Table 2). This was complemented by moderate correlations with *Average.long.pass.length.m*, *Passes.to.penalty.area.per.90*, and *Passes.to.-final.third.per.90*.

FBref showed similar patterns, with *Long.passes.per.90* achieving a correlation of 0.638, while *Crosses.into.penalty.area* (0.558) also showed increased correlation compared to their relationship with *Total OBV*. The dataset maintained at least five variables with correlations exceeding 0.4, mirroring the strength of relationships found in Wyscout.

The importance of long passes and penalty area entries in both datasets suggested that linear modeling methods could effectively capture *Pass OBV*, regardless of which dataset was used as the foundation for a scouting model.

4.4.1.2 Pass OBV PCA

When applying PCA to the *Pass OBV* data, both Wyscout and FBref datasets' scree plots (Figures A8 A and B in the Appendix) indicated that many PCs were needed for adequate variance explanation, implying potential limitations of linear dimension reduction techniques. Figure 18 below shows the biplot of the first two PCs on these *Pass OBV* datasets, with a *Pass OBV* heatmap underneath it and the five different playing positions highlighted.

The first two PCs in both datasets showed similar patterns in player position separation, with notable overlapping between midfielders and fullbacks, while center backs and forwards tended to be more isolated.

The first two PC loadings are shown in the appendix in Table A9 (Wyscout) and Table A10 (FBref). Wyscout's PC1 reflected attacking creativity and goal involvement. High positive loadings for variables like *xA.per.90* (0.358), *Smart.passes.per.90* (0.288), and *Passes.to.penalty.area.per.90* (0.325) highlighted a focus on chance creation, key passes, and penetrating defenses. Negative loadings for variables like *Average.pass.length.m* (-0.299) and *Accurate.long.pass.length.m* (-0.275) suggested less emphasis on conservative, deep-lying passing, reinforcing PC1 as a measure of creative, goal-focused playmaking. Unlike Wyscout's PC1, which emphasized attacking creativity, FBref's PC1 reflected passing volume and efficiency in structured buildup play. High positive loadings for *Progressive.passing.distance.per.90* (0.389), *Passes.into.final.3rd.per.90* (0.369), and *Accurate.short.passes.percent* (0.351) highlighted frequent, precise progression and control, typically associated with defenders (as is seen with the high loadings of PC1 for center backs in Figure 18 above). In contrast, negative loadings for *xA.per.90* (-0.143) and *Passes.into.penalty.area.per.90* (-0.055) suggested more focus on high-risk, creative actions, positioning this PC1 as a measure of reliable and controlled passing for maintaining possession and advancing

play.

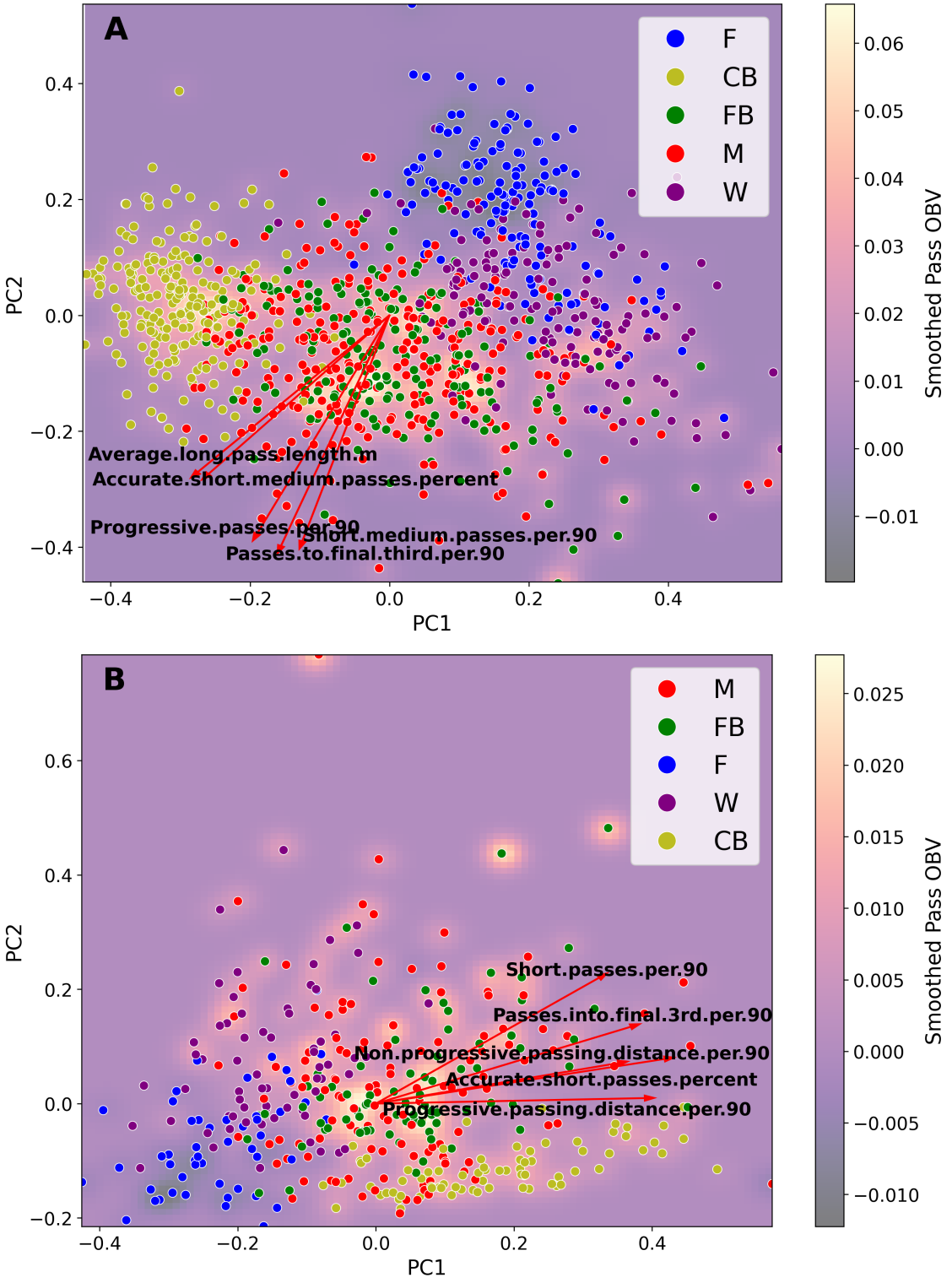

Figure 18: PCA Biplots of Wyscout (A) and FBref (B) passing datasets

The second principal component revealed complementary insights across both datasets. Wyscout's PC2 was characterized by predominantly negative loadings, with the strongest negative contributions from *Passes.to.final.third.per.90* (-0.398), *Short.medium.passes.per.90* (-0.390), and *Progressive.passes.per.90* (-0.377). The consistent negative loadings across various passing metrics suggested that this component might have distinguished players based on their tendency to avoid certain types of passes, particularly those involving progression up the field. This was inline with what we expected as Forwards were seen to have large loadings on this second PC, indiciating their lack of involvement in passing the ball up the pitch. FBref's PC2 (Table A10) primarily captured creative and attacking passing ability, with the strongest positive loadings found in *Passes.into.penalty.area.per.90* (0.489), *Expected.assists.per.90* (0.468), and *Progressive.passes.to.final.third.per.90* (0.347). Additional significant contributions came from *Through.balls.per.90* (0.343), *Crosses.into.penalty.area.per.90* (0.330), and *Assists.per.90* (0.317), further emphasizing this component's focus on offensive passing. These loadings collectively indicated that PC2 represented a player's ability to make penetrative, attacking passes that directly contributed to chance creation. This was further shown by the heatmap yielding high values (lighter colors) for individuals with high PC2 loadings.

4.4.1.3 *Pass OBV* t-SNE

The t-SNE *Pass OBV* plots in Figure 19 below showed that both Wyscout and FBref datasets arguably demonstrated superior clustering capabilities compared to PCA, especially in position-based clustering. The Wyscout analysis (Figures 19 A and B) positioned defenders on the left and attackers on the right, while FBref (Figures 19 C and D) placed defenders in the middle and top-right corner, with attackers on the left.

A key finding across both datasets was the distinct positioning of fullbacks. In both cases, fullbacks with high *Pass OBV* values formed a separate cluster, distinctly separated from center backs, forwards, and wingers. This indicated t-SNE's ability to identify unique passing characteristics of fullbacks that differentiated them from these other positions. This separation was notably clearer than in the PCA visualizations, with both datasets showing significantly reduced positional overlap.

FBref's analysis additionally identified multiple high-performing clusters, particularly a mixed cluster of fullbacks, wingers, and a midfielder (bottom of Figure 19 D), implying that similar passing styles transcended traditional positional categories. These findings, particularly pronounced in the FBref dataset, strengthened the case for using non-linear dimension reduction methods to model *Pass OBV*, as these methods appeared better suited to capturing the nuanced relationships between passing styles and performance.

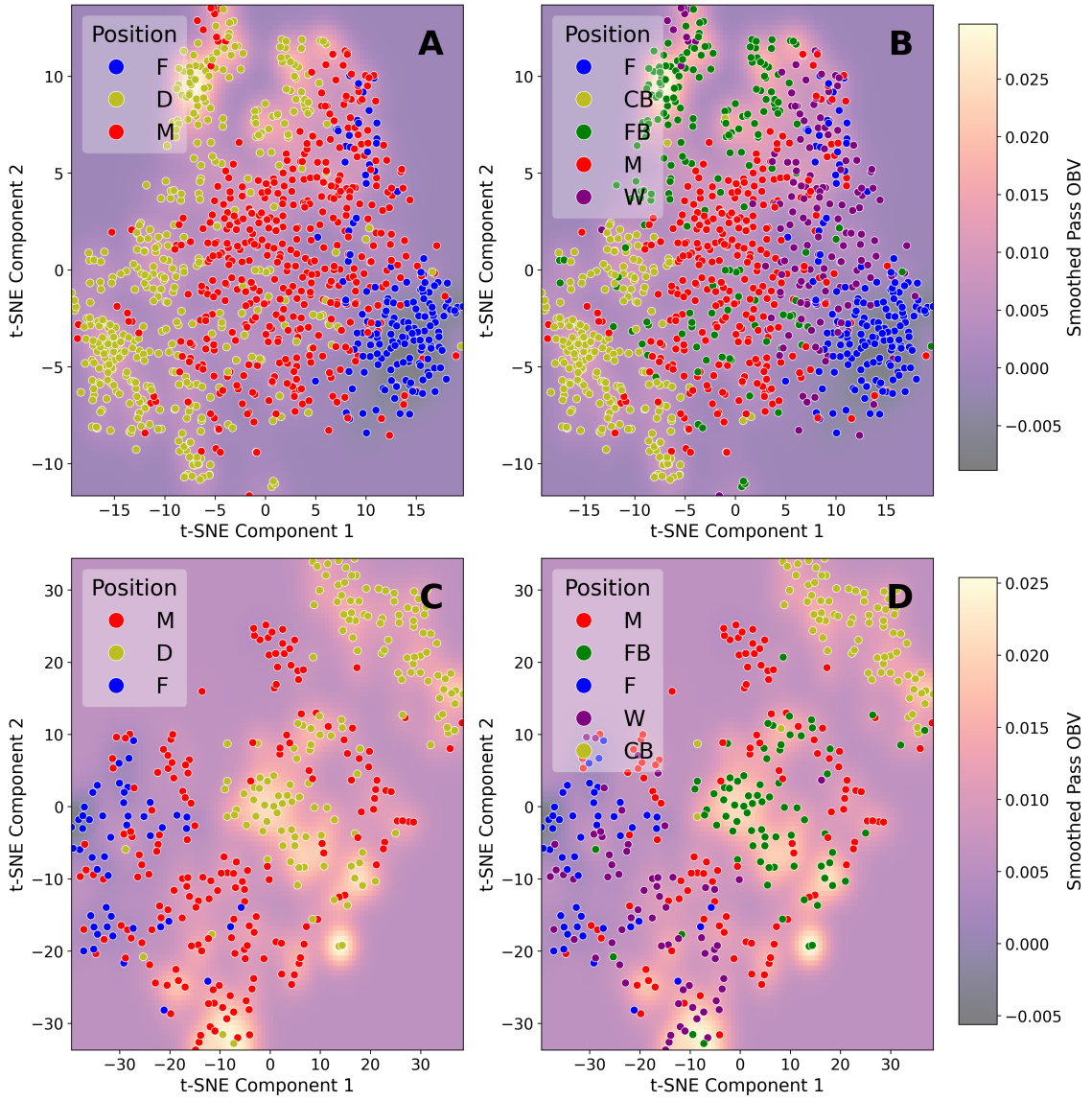

Figure 19: First two t-SNE components from Wyscout (A, B) and FBref (C,D) passing data overlaid onto a *pass OBV* heatmap

4.4.1.4 Pass OBV Random Forest Variable Importance

To further understand which variables may have held the greatest predictive power when modeling *Pass OBV*, a Random Forest variable importance plot was computed for the Wyscout (Figure 20 A) and FBref (Figure 20 B) datasets.

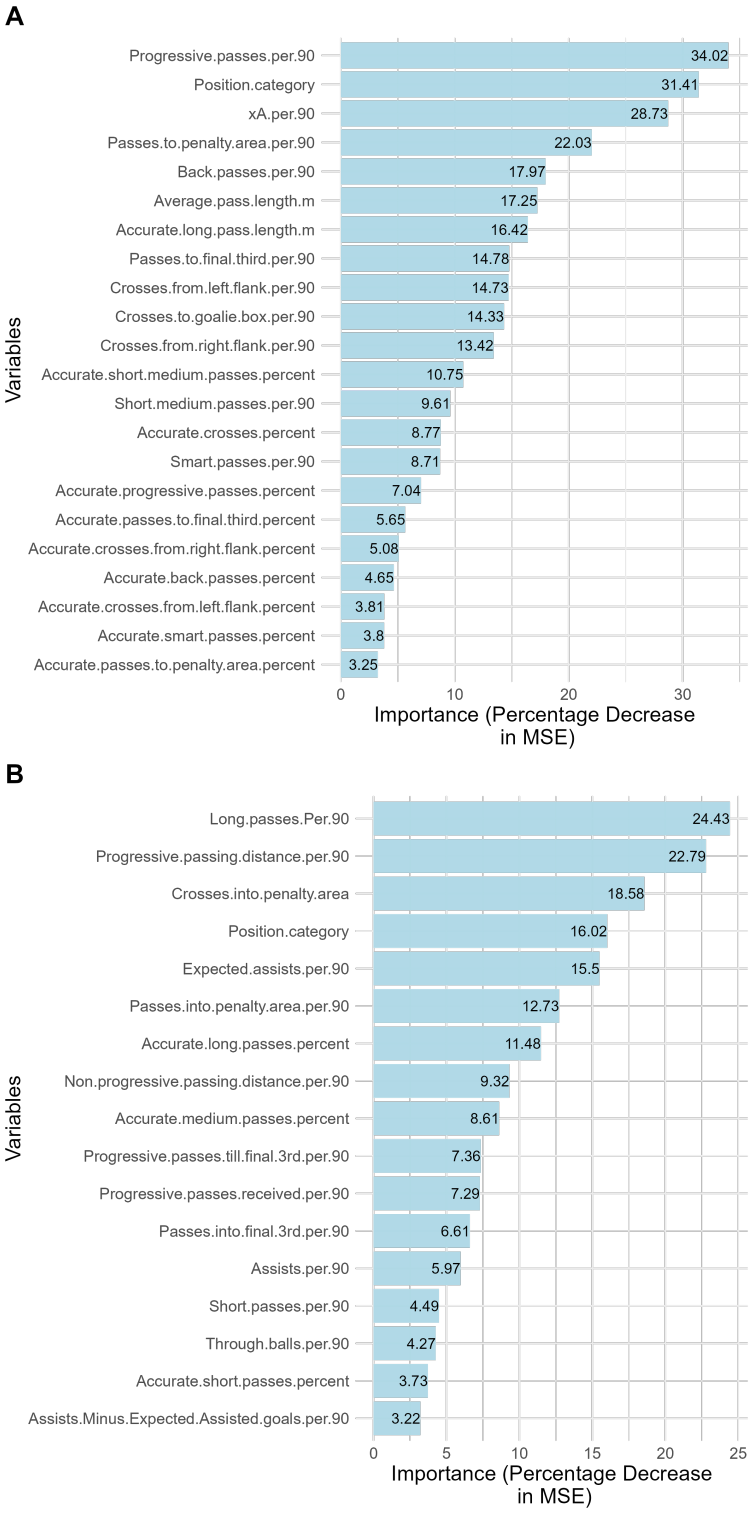

Figure 20: Random forest variable importance plot for Wyscout (A) and FBref (B) datasets

Both Wyscout and FBref *Pass OBV* datasets revealed similar patterns in their most influential predictors, with some notable differences in specific metrics. In Wyscout's data, *Progressive.passes.per.90* emerged as the leading predictor, closely followed by *xA.per.90*. In the FBref data, *Long.passes.per.90* and *Progressive.passing.distance.-per.90* were identified as its most important variables. Despite these differences in specific metrics, both analyses highlighted the importance of forward-moving, attacking passes. This was in line with expectations, as shown above (Figure 5), where attacking passes garnered the greatest *Pass OBV* values. Player position emerged as a significant predictor in both datasets, ranking highly in importance and confirming that positional context was crucial for evaluating passing contribution. This finding supported the need to consider player positions when modeling *Pass OBV*.

Interestingly, both analyses showed that frequency-based metrics generally outweighed accuracy-based ones. In Wyscout, pass accuracy metrics across various types (long, medium, short, progressive, crosses) ranked lower than frequency metrics, while FBref similarly showed that metrics like assists, through balls, and short pass accuracy had less impact than the volume of long and progressive passes. This finding indicated that the quantity and ambition of passing attempts might be more valuable for predicting *Pass OBV* than pure accuracy. This could be considered a shortcoming of the metric, as players who frequently attempted difficult passes (such as long, hopeful passes up the field) could accumulate higher *Pass OBV* scores despite lower completion rates, potentially overshadowing those who executed simpler passes with greater consistency.

4.4.1.5 Pass OBV Potential Interaction Terms

Given that position appeared to be a crucial component of *OBV* and that the utility of linear models to predict *Pass OBV* was a key research question, position-based interaction terms were investigated when modeling *Pass OBV*. The results of this analysis are displayed in Figure 21 below.

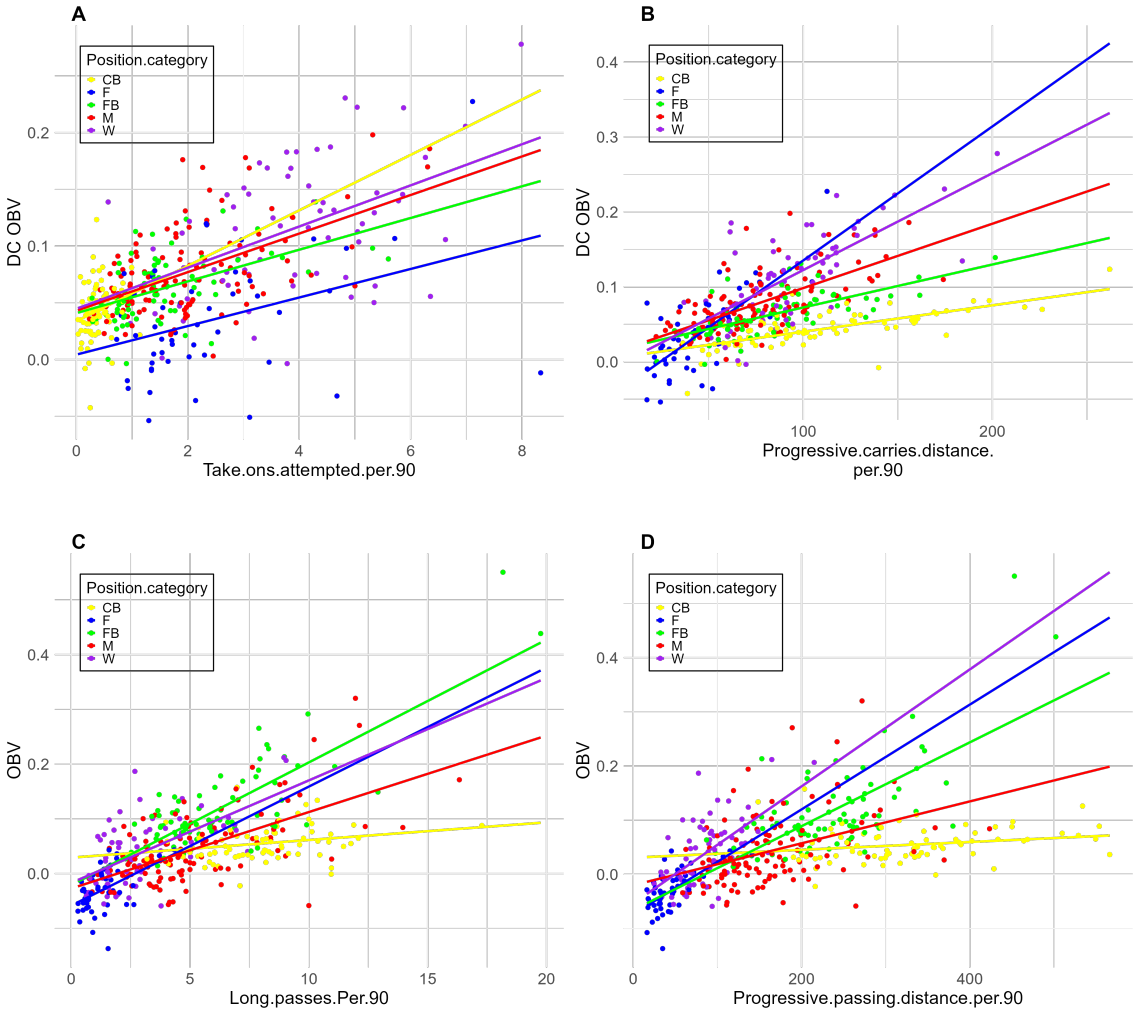

Figure 21: Wyscout (A,B) and FBref (C,D) passing variables' interaction with player position

In both datasets, forwards and fullbacks gained the most *OBV* from progressive actions—forwards due to their advanced positions, making dangerous passes near goal, and fullbacks through their ability to deliver high-value crosses and diagonal balls. Both datasets also showed center backs and midfielders receiving less *OBV* per progressive action, likely due to their deeper positioning. This positional influence was particularly evident in long passes, where forwards and fullbacks again showed the highest *OBV* returns, while the relationship between progressive distance/frequency and *OBV* consistently favored players in more advanced positions. These findings, as hypothesized previously, implied that position-based interaction terms would be

critical in linear models.

4.4.2 Dribble and Carries (DC) OBV

4.4.2.1 DC OBV Correlation Analysis

The Wyscout dribbling dataset offered just four dribbling variables (Table 4 below), with *Progressive.runs.per.90* showing the strongest correlation (0.619), followed by *Accelerations.per.90* (0.537). The percentage of successful dribbles had a very weak correlation, once again suggesting that frequency mattered more than success rate. This limited variable set raised concerns about the adequacy of modeling, particularly linear modeling, for Wyscout's *DC OBV* prediction.

FBref presented a more comprehensive dribbling dataset, with five variables exceeding 0.5 correlation, and notable features including some locational information. *Carries.into.final.3rd.per.90* emerged as the strongest correlate (0.68), followed by *Carries.into.penalty.area.per.90* (0.588) and *Take.ons.attempted.per.90* (0.568). This richer set of metrics implied that FBref would be offering greater linear insights into *DC OBV* than Wyscout's dataset. This was largely due to its richer locational information, as the previous Figure 6 showed that *DC OBV* was largely affected by where the dribbling actions took place.

Wyscout Dataset		**FBref Dataset**	
Variable	**Correlation**	**Variable**	**Correlation**
Progressive.runs .per.90	0.619	Carries.into .final.3rd.per.90	0.680
Accelerations .per.90	0.537	Carries.into .penalty.area.per.90	0.588
Dribbles .per.90	0.485	Take.ons .attempted.per.90	0.568
Successful.dribbles .percent	0.129	Non.progressive .carries.distance.per.90	0.522
		Progressive.carries .distance.per.90	0.504

Table 4: Top variable correlations with *DC OBV* in the Wyscout and FBref datasets

4.4.2.2 DC OBV PCA

The datasets showed markedly different PCA characteristics in their dribbling analysis scree plots (Figure A9). Wyscout's scree plot revealed that just two components accounted for roughly 90% of variation, with the first PC alone capturing over

60% through positive loadings from three highly correlated variables. However, this apparently efficient dimensional reduction failed to translate into meaningful player separation (Figure 18 in the Appendix), showing significant positional overlap. Only wingers achieved partial segregation with high first PC loadings, while fullbacks spread widely across the second PC. These results confirmed what was suspected about the challenges of modeling *DC OBV* using Wyscout's dribbling metrics, as players across different positions appeared to exhibit similar dribbling patterns. The lack of clear groupings among players who excelled in this area indicated that the dataset was lacking more granular variables.

FBref presented a more complex picture, requiring four components to reach 90% variance explanation (Figure A9), although the first two components explained a substantial 75%. The first PC represented overall carrying intensity across various metrics, while the second PC was more nuanced, contrasting safer ball carrying (positive loadings for *Carries.per.90* and *Progressive.carrying.distance.per.90*) against riskier, high-impact actions (negative loadings for *Carries.into.penalty.area.per.90* and *Take.on.attempted.per.90*). The PCA revealed some clusters of high-performing players, particularly those with negative scores on the first PC. This suggested that the most valuable dribbling actions in terms of OBV likely occurred in positions that were very advanced up the field, rather than just carrying the ball into the final third of the field, which was what Table 4 might have suggested.

The key distinction between the two datasets lay in their effectiveness: Wyscout's PCA failed to differentiate positions and styles effectively, suggesting limited utility for linear modeling. In contrast, FBref's analysis identified distinct clusters of high-performing dribblers, indicating potential value for PCA in modeling *DC OBV*, despite some positional overlap. This difference likely stemmed from FBref's more comprehensive dribbling metrics, allowing for better capture of stylistic nuances in play.

4.4.2.3 DC OBV t-SNE

The effectiveness of t-SNE again differed notably between the datasets. Wyscout's analysis (Figure 25 A and Figure 25 B) struggled to detect meaningful stylistic patterns between the positions, implying that it had not captured position-related characteristics in dribbling, which is essential when modeling OBV. The only significant finding was a cluster of high *DC OBV*-scoring wingers in the bottom right corner. The algorithm otherwise failed to provide meaningful separation or insights.

FBref's analysis also found this high-performing cluster of wingers, achieving semi-distinct positional separation. This suggested that the position-related variables in the FBref dataset had allowed the algorithm to better identify which players were

more involved in certain actions, depending on their position. While Wyscout's limited variables constrained even non-linear analysis, FBref's results suggested potential value in applying non-linear dimension reduction methods to aid in the modeling of *DC OBV*.

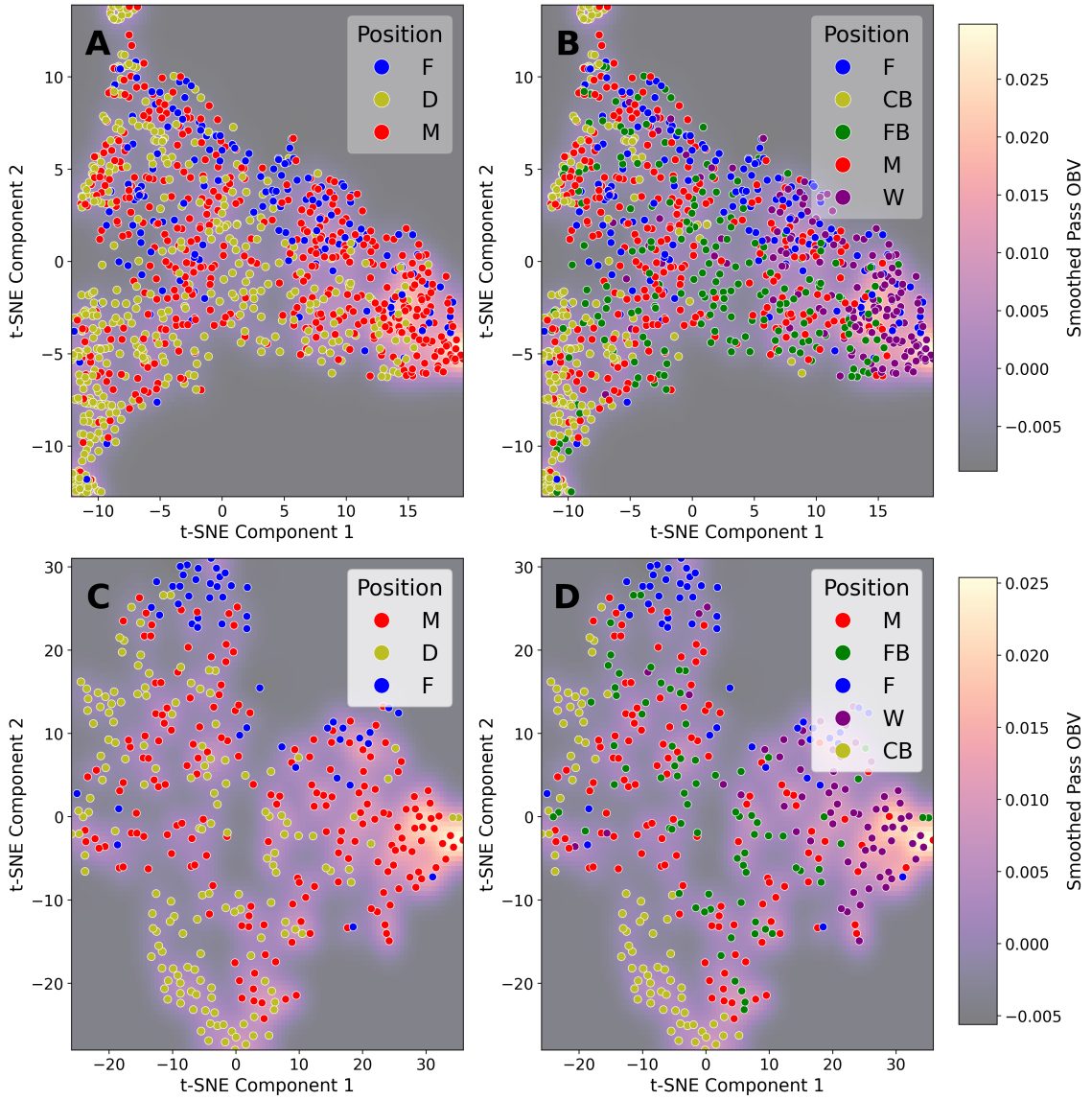

Figure 22: First two t-SNE components from Wyscout (A, B) and FBref (C,D) dribbling data overlaid onto an*DC OBV* heatmap

4.4.2.4 DC OBV Random Forest Variable Importance

The results of applying a Random Forest variable importance analysis to both dribbling datasets are displayed in Figure 23 below. Both datasets identified player position as a crucial predictor, though with different relative importance. In the Wyscout data (23 A), position emerged as the most critical feature, while in the FBref dataset (23 B), it ranked as important but was secondary to specific carrying metrics. This suggested that positional context significantly influenced dribbling value across both datasets, though it was captured differently through their respective variables. The model's over-reliance on the *Position.category* variable in the Wyscout dataset was a concern. With a variable importance value of 45, it suggested that nearly half of the model's predictive accuracy was attributed to a player's position. While position was undoubtedly important, this heavy reliance likely indicated a significant misinterpretation of *DC OBV*, stemming from the limitations and sparsity of the Wyscout dataset available to the model.

The datasets then diverged in their key performance indicators. In the Wyscout dataset, *Progressive.runs.per.90* emerged as the second most important predictor, followed by *Dribbles.per.90*. In contrast, when using the FBref dataset, *Carries.into.final.3rd.per.90* was the primary predictor, followed by *Progressive.carries.distance.per.90* and *Carries.into.penalty.area.per.90*. This difference indicated that the FBref dataset was more robust, incorporating more positional information, such as where the dribbles were taking place.

Both analyses showed consistency in downplaying certain metrics. In the Wyscout dataset, the percentage of successful dribbles was ranked as the least important among the variables, as suggested in Table 4 above, while the FBref dataset similarly assigned lower importance to non-progressive carrying distance despite its high correlation with *DC OBV*. This implied that the random forest model applied to the FBref data identified more complex relationships between variables than those indicated by simple correlations.

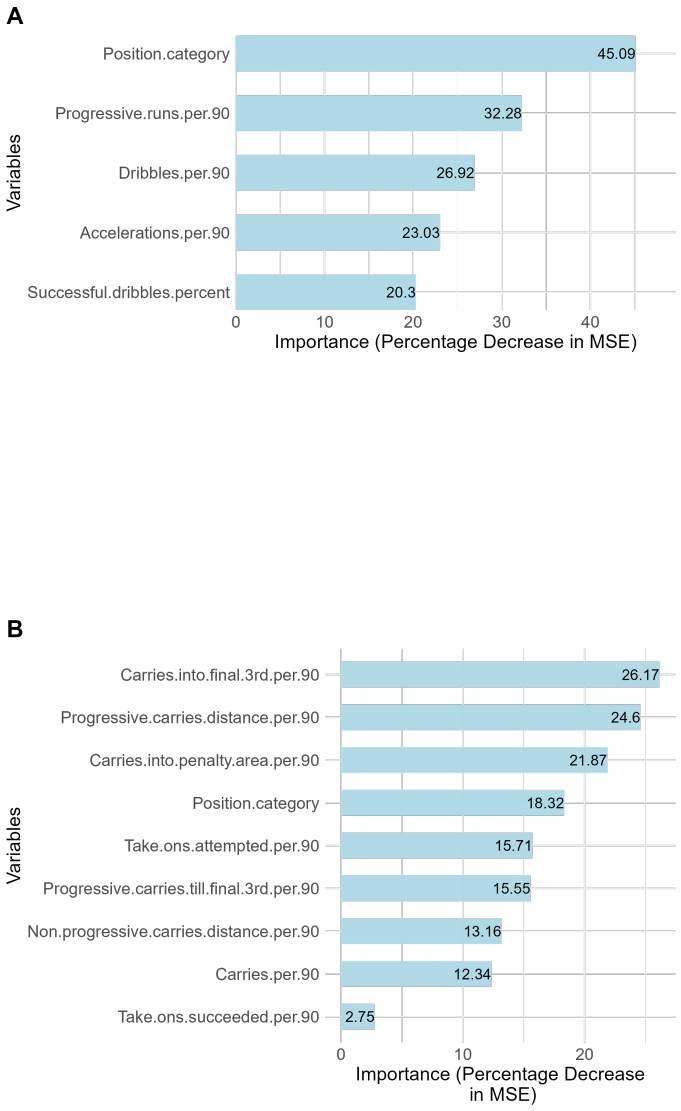

Figure 23: Random Forest variable importance with Wyscout(A) and FBref(B) dribbling variables modelled on *DC OBV*

4.4.2.5 DC OBV Potential Interaction Terms

Interactions between the dribbling variables that demonstrated the most significant differences in their relationships with position and *DC OBV* were examined. Unsurprisingly, the dribbling variables that naturally aligned with interaction terms involving *Position.category* were those that lacked explicit locational information. This was likely because *Position.category* served as a proxy for a player's location on the field, allowing the model to better interpret these variables through their interaction with position. In contrast, locational variables such as *Dribbles.into.the.final.3rd.per.90* did not require interaction with *Position.category*, as the location of the action (the final third) was already inherent in the variable itself.

Figure 24 below presents the relationship between *Progressive.runs.per.90*, *Dribbles.per.90*, and *DC OBV* from the Wyscout dataset (Figure 24 A and Figure 24 B), and *Take.ons.attempted.per.90*, *Progressive.carries.distance.per.90*, and *DC OBV* from the FBref dataset (Figure 24 C and Figure 24 D). Each figure contains a line of best fit for different positions, allowing for an easy understanding of the effect of a unit change in one of the variables on *DC OBV*.

Plot A revealed that both the intercepts and slopes varied across positions, indicating that the effect of a progressive run on *OBV* differed by position. This variation suggested that an interaction term between position and the number of progressive runs would be important for improving the accuracy of a linear model. Similarly, for *Dribbles.per.90* (Plot B), the data showed that center backs experienced the greatest increase in *OBV* for each additional dribble, while midfielders exhibited the least increase in *DC OBV* for an additional dribble. These observations indicated that including interaction terms in the model would enhance its performance in capturing the nuanced effects of these variables across different positions.

When performing a similar analysis on the FBref dataset (Figure 24 C and D below), center backs exhibited the highest *DC OBV* return for an additional take-on attempted, whereas the rest of the positions had more similar returns for such an action. However, when turning attention to the distance a player progressively carried the ball, an intuitive finding previously discussed was substantiated. Plot D demonstrated that the amount of *DC OBV* attributed to each unit of distance a player progressively carried the ball varied significantly by position. Forwards received the highest *DC OBV* for a unit change in distance carried, followed by wingers, midfielders, fullbacks, and finally center backs. This pattern suggested that the further up the pitch a player's general position was, the more *DC OBV* was typically assigned to their progressive carries. This observation aligned with what was shown in Figure 6 earlier, indicating that each unit of distance covered progressively with

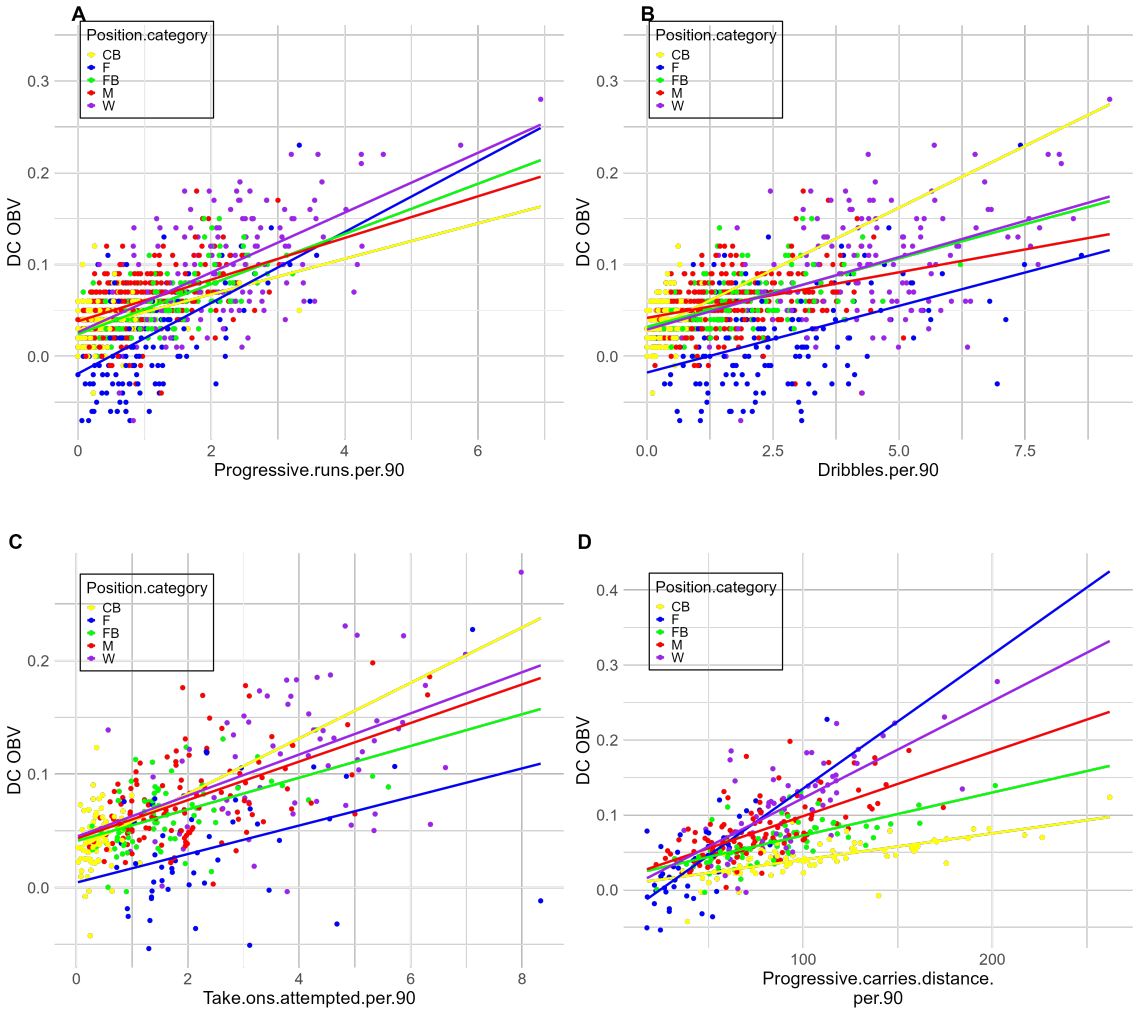

Figure 24: Wyscout (A,B) and FBref (C,D) dribbling variables' interaction with position

the ball was more valuable in forward areas than in defensive zones. This further implied that linear models would benefit from interaction terms.

4.4.3 Defensive Actions (DA) OBV

4.4.3.1 DA Correlation analysis

Table 5 below shows the top five correlations with *DA OBV* from the two datasets.

Wyscout Dataset		FBref Dataset	
Variable	**Correlation**	**Variable**	**Correlation**
Shots.blocked .per.90	0.333	Interceptions .per.90	0.381
PAdj.Interceptions	0.311	Tackles.in.defensive .3rd.per.90	0.252
Defensive.duels.won .percent	0.257	Tackles.in.midfield .3rd.per.90	0.146
Aerial.duels.won .percent	0.195	Tackles.in.attacking .3rd.per.90	0.062
		Yellow.cards .per.90	0.037

Table 5: Top 5 variable correlations with *DA OBV* in the Wyscout and FBref datasets

Both datasets exhibited notably low correlations between defensive metrics and *DA OBV*. *Interceptions.per.90* were present in both datasets, while the Wyscout dataset additionally highlighted *Shots.blocked.per.90* (0.333) as its strongest correlator. From the exploratory data analysis (Figure 7), it was observed that defensive actions in the penalty boxes held the highest value, with this value dropping sharply outside these regions. However, neither dataset included "in-box" variables, making it challenging to effectively model *DA OBV*. While the FBref dataset provided *Tackles.in.attacking.third.per.90* and *Tackles.in.defensive.third.per.90*, these broad zones encompassed both high- and low-value areas, limiting their precision. Furthermore, the consistently weak linear relationships across both datasets indicated that non-linear modeling approaches may perform better when it comes to modeling *DA OBV*.

4.4.3.2 DA PCA

Both datasets revealed limitations in effectively reducing the defensive datasets using PCA. Figure A12 A in the Appendix demonstrated that the Wyscout PCA failed to uncover meaningful positional patterns, with the biplot revealing no clear differences in tackling habits across positions. The only positive finding was that high-performing *DA OBV* players generally had negative loadings on the first PC, indicating higher values for *Duels.won.percent*, *Shots.blocked.per.90*, and *Defensive.duels.won.percent*. However, mid-level performers were still scattered among these high performers. Similarly, the FBref analysis (Figure A12 B in the Appendix) showed minimal positional clustering, with the only notable observation being that defend-

ers tended to have the lowest loadings on the second PC, reflecting their minimal involvement in tackles in the attacking third. This result underscored the need for more granular locational data to better differentiate player profiles. These findings from both datasets suggested that PCA would be suboptimal when using it as a dimension reduction technique before modeling *DA OBV*, and pointed towards non-linear approaches as being potentially more effective.

4.4.3.3 DA OBV t-SNE

Subplots A and B in Figure 25 show distinct separation of players based on positions in the Wyscout dataset, with center backs being the sole high-performing cluster in these plots. Subplots C and D show that the FBref allow for even more granular isolation of high-performing defenders than that achieved on the Wyscout dataset, with a clear separation between center backs, fullbacks, and the rest of the players.

The t-SNE analyses performed on each dataset further highlighted the challenges in modeling *DA OBV*, though with some differences. t-SNE performed on the Wyscout dataset successfully clustered high-performing center backs and showed clear positional separation but failed to identify high-performing clusters for other positions. The FBref dataset captured greater complexity, revealing more distinct clusters but showing difficulty in distinguishing between defensive positions that had similar *DA OBV* values. These results further highlighted that predicting *DA OBV* would be particularly challenging due to factors not captured in the available variables.

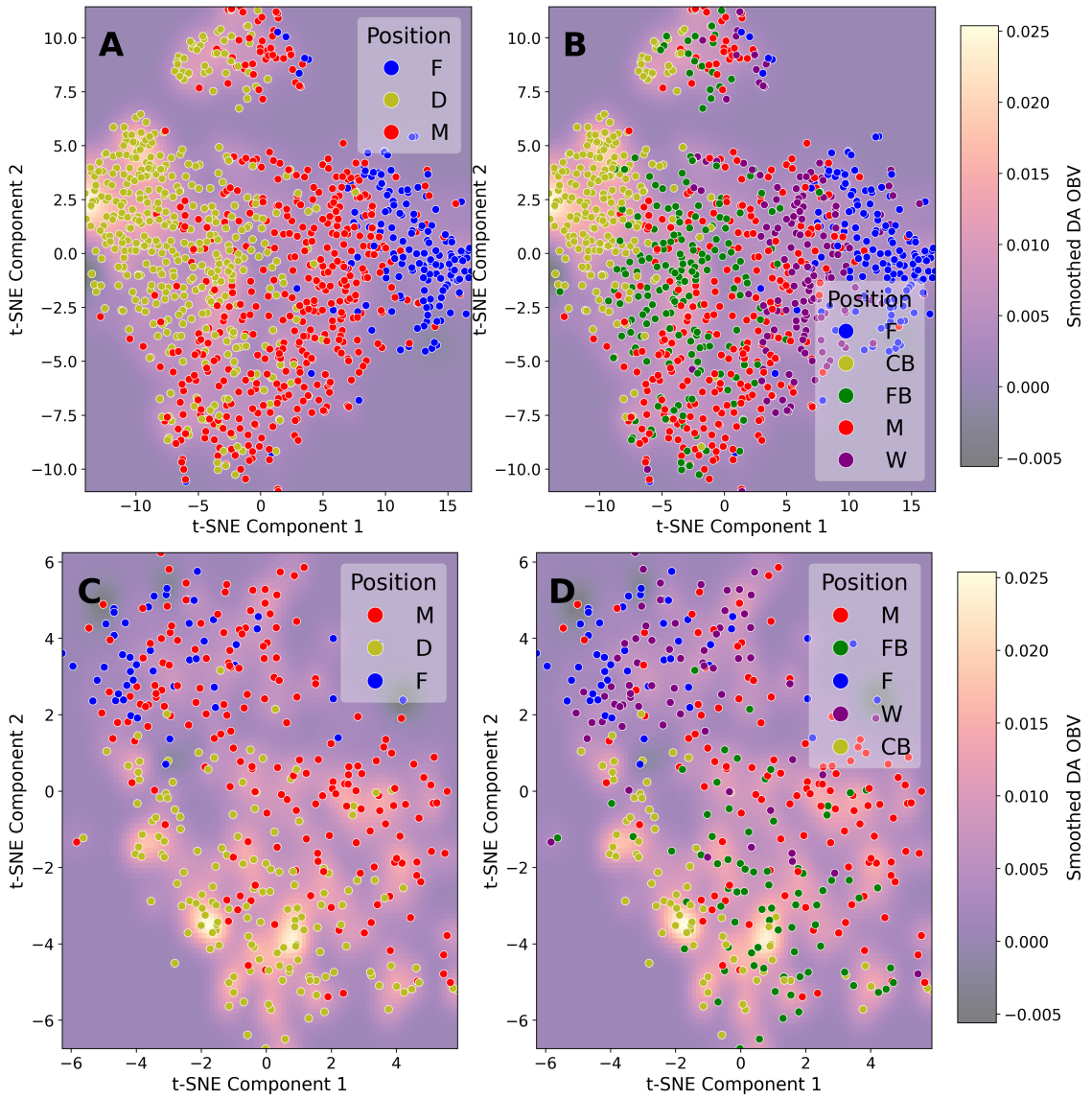

Figure 25: First two t-SNE components from Wyscout (A, B) and FBref (C,D) defensive data overlaid onto a *DA OBV* heatmap

4.4.3.4 *DA OBV* Random Forest Variable Importance

Figure 26 shows that both Random Forest analyses revealed interesting contrasts in variable importance for predicting *DA OBV*. In the Wyscout data, *Duels.won.percent* emerged as the top predictor, but with a relatively modest importance (16% MSE increase if removed). This suggested that *DA OBV* prediction

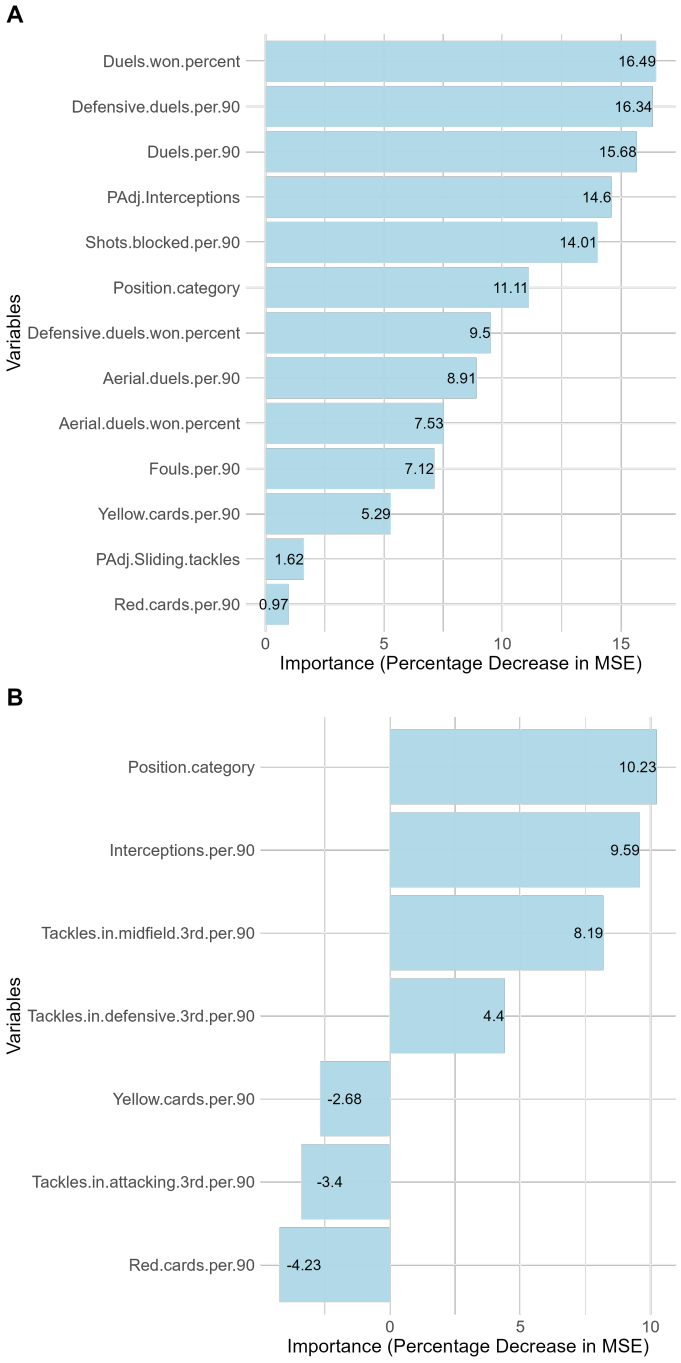

Figure 26: Random Forest variable importance with Wyscout(A) and FBref(B) defensive variables modeled on *DA OBV*

relied on a combination of variables rather than a single dominant metric. In the

FBref data, position was the primary predictor, followed by *Interceptions.per.90* and Tackles in different pitch zones.

A key difference emerged in the importance of playing position between the two datasets: while it was one of several important factors in the Wyscout data, it dominated the importance ranking in the FBref data. An important insight was that both analyses suggested challenges in effectively modeling *DA OBV*—the distributed variable importance in the Wyscout data potentially indicated a lack of clear patterns, while the FBref data showed a potential over-reliance on positional information.

4.4.3.5 *DA OBV* Potential Interaction Terms

Figure 27 below highlights patterns in both datasets regarding how defensive actions relate to *DA OBV* with respect to position. In the Wyscout data (subplots A and B), *PAdj.interceptions.per.90* showed relatively consistent patterns across positions, while *Shots.blocked.per.90* demonstrated slightly more positional variation. Notably, these relationships exhibited non-linear characteristics, and as such, linear models did not accurately capture these patterns.

The analysis of the FBref dataset, focusing on *Interceptions.per.90*, showed that center backs and forwards achieved the highest *DA OBV* per interception. This finding aligned with what was seen in Figure 7, as center backs are often positioned in their own box—a high-value area for defensive actions—while forwards are further advanced, where the *DA OBV* value attributed to defensive actions is also high. Both analyses suggested that including interaction terms could aid in linear modeling. However, given the strong non-linear patterns observed in these relationships, even enhanced linear models would likely struggle to fully capture a significant amount of variation. The absence of locational data remained a significant limitation, implying that even sophisticated non-linear models might have lacked important information needed to accurately model *DA OBV*.

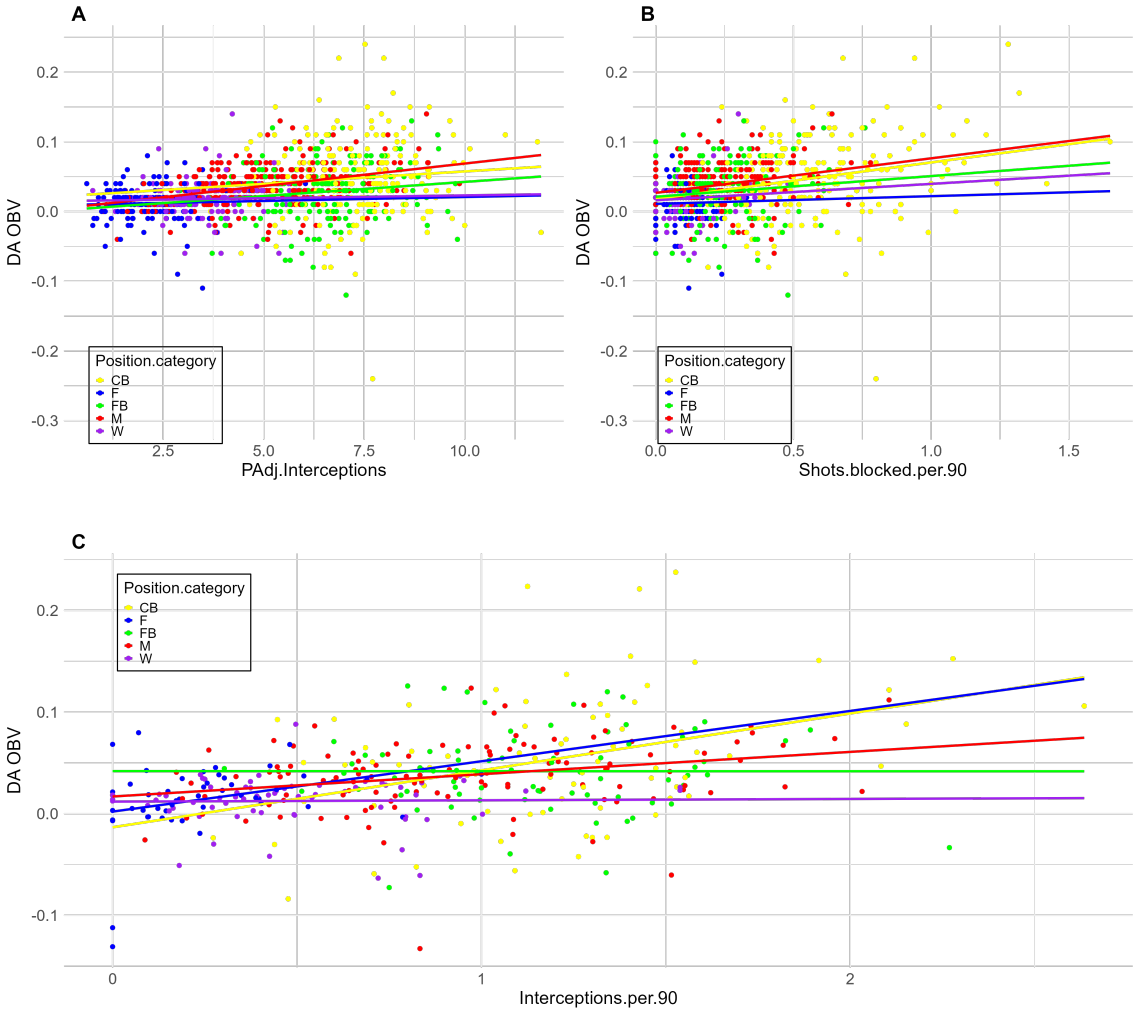

Figure 27: Wyscout (A,B) and FBref (C) defensive variables' interaction with position

4.5 Conclusion

This EDA sought to determine the feasibility of modeling the complex, location-dependent *OBV* metric using more accessible, generally non-location-based frequency statistics from the Wyscout and FBref datasets. The investigation was guided by three core interests: the validity of *OBV* as a performance metric, the strength of its relationship with the available datasets, and whether to model *OBV* holistically or by its individual components.

The analysis first validated *OBV* as a meaningful measure of team quality. A strong

positive correlation was found between a team's cumulative *OBV* rank and its final league position (Spearman coefficients of 0.72 and 0.75), confirming that teams with higher *OBV* are indeed more successful. This finding underscored the value of *OBV* as a target variable for a scouting departments. The feasibility of modeling *OBV* using non-locational data proved to be nuanced and highly dependent on the specific component being analyzed. The analysis strongly supports modeling *OBV* through its individual components—*Pass OBV*, *DC OBV*, and *DA OBV*—rather than as a single entity. Each component exhibited distinct distributional properties and relationships with the explanatory variables. *Pass OBV* showed the most promise for modeling, with both datasets containing variables that correlated strongly with it, particularly those related to progressive passing and entries into the penalty area. In contrast, modeling *DA OBV* was identified as the most significant challenge due to consistently weak correlations across both datasets. This limitation stemmed from the inability of the available frequency-based statistics to capture the critical locational context that determines the value of defensive actions, such as an interception in the penalty box versus one in the midfield. The potential for modeling *DC OBV* was moderate, with the richer, quasi-locational metrics in the FBref dataset offering a more robust foundation than the sparser Wyscout data.

A central theme that emerged was the superiority of non-linear methods for understanding the data's underlying structure. The consistent success of t-SNE in clustering players by position and performance—compared to the limited efficacy of PCA—demonstrated that the relationships between frequency statistics and *OBV* are inherently non-linear. While linear models were deemed potentially viable for *Pass OBV* and, to a lesser extent, *DC OBV*, their success would be contingent on careful feature engineering, most notably the inclusion of position-based interaction terms. The interaction analysis confirmed that the value of an action (e.g., a progressive run) varies significantly based on a player's position, which serves as a crucial proxy for on-field location.

This EDA provides compelling evidence that while modeling *OBV* is challenging, it is a worthwhile endeavor. A component-based modeling approach is recommended, acknowledging that the predictability of each component varies. Non-linear models appear best suited to capture the complexity of player actions, though enhanced linear models may suffice, specifically for the passing component. Furthermore, the findings establish a clear path forward for the modeling phase, highlighting the importance of player position, the distinct nature of each *OBV* component, and the inherent limitations of using non-locational data to predict a spatially-driven metric.

5 Methods

To test the hypothesis that OBV can be modeled using readily available frequency statistics, enabling teams of all financial capacities to obtain OBV estimates across multiple leagues at a lower cost, a comprehensive analysis was conducted. This included both simple and advanced data exploration and modeling techniques to evaluate the feasibility and effectiveness of this approach. This section outlines the procedures applied to the data to first understand it, facilitating more informed advancements in modeling, followed by a discussion of the modeling techniques used and how they were implemented. Exploratory methods included Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Random Forests for assessing variable importance. Dimension reduction techniques applied for modeling comprised PCA, Autoencoders and using measures of variable importance to dictate which variables to include. The predictive models employed were Linear Regression, Multi-layer Perceptron (MLP), Random Forests, and Extreme Gradient Boosted Trees.

5.1 Dimension Reduction and Exploratory Models

5.1.1 PCA

This section is based on the paper "*Principal component analysis*" by Wold et al. (1987).

5.1.1.1 Introduction

Principal Component Analysis (PCA) is a powerful statistical method used to reduce the number of dimensions in high-dimensional datasets. It transforms a large set of variables into a smaller set through linear combinations of the variables with the aim of explaining a significant portion of the variance in the data. By projecting data onto a new coordinate plane, PCA identifies the directions (principal components) along which the variation in the data is maximized. In this way it simplifies the complexity of the dataset while attempting to preserve its structure and relationships.

5.1.1.2 Underlying Assumptions

PCA operates under several key assumptions:

- **Linearity:** The relationships between variables are linear. If non-linear relationships are present, the reduction process may be sub-optimal.

- **Large variance equals importance:** Variables with higher variance are more significant and carry more information, thus PCA is sensitive to the scale of features. For this reason, we will scale the features before applying PCA to them.
- **Orthogonality of principal components:** The principal components are orthogonal to each other, ensuring that they are uncorrelated and independent. Resultant features are thus uncorrelated, which can be useful for modeling purposes but may not capture the true underlying distribution of the data.

5.1.1.3 Algorithm and Mechanism

The PCA algorithm involves the following steps, assuming there are p variables, and n observations:

1. **Standardization**: The data is standardized (mean-centered and scaled to unit variance) to ensure that each variable contributes equally to the analysis, as opposed to the analysis being biased to variables with larger scale. Equation x below shows the transformation of a data point.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{5.1.1.1}$$

   where z_{ij} is the transformed value for the i-th observation of the j-th variable, and x_{ij} is the original value. $\bar{x}_j$ is the mean of the j-th variable across all observations, defined as:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{5.1.1.2}$$

   where n is the total number of observations. Finally, s_j is the standard deviation of the j-th variable, calculated as:

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} \tag{5.1.1.3}$$

   The standardized value z_{ij} represents how many standard deviations the original value x_{ij} is away from the mean $\bar{x}_j$.

2. **Covariance Matrix Computation**: The covariance matrix is then calculated to understand how variables in the dataset relate to each other.

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \tag{5.1.1.4}$$

where $\mathbf{X}_i$ is a $p \times 1$ column vector representing the i-th observation, and $\bar{\mathbf{X}}$ is the mean vector of size $p \times 1$. Therefore, $\mathbf{C}$ is a covariance matrix of dimensions $p \times p$, containing the covariances between the p variables.

3. **Eigenvalue Decomposition** Eigenvalue Decomposition is then performed on the covariance matrix $\mathbf{C}$ to identify the eigenvalues and corresponding eigenvectors which maximise the explained variance of the full dataset. The eigenvectors, denoted by the matrix $\mathbf{V}$, indicate the directions of the principal components on the Cartesian plane, while the eigenvalues, represented by the diagonal matrix $\mathbf{\Lambda}$, indicate the magnitude of the variance explained by each component. Each eigenvalue thus corresponds to a single principal component, and quantifies the associated variance explained by that component. The general eigenvalue equation for a covariance matrix $\mathbf{C}$ is expressed as:

$$\mathbf{C}\mathbf{v_i} = \lambda_i \mathbf{v_i} \tag{5.1.1.5}$$

where $\mathbf{v}_i$ is the $p \times 1$ eigenvector corresponding to the i-th principal component, and λ_i is the eigenvalue indicating the variance explained by this component. To obtain these values, the following equation is solved:

$$\det(\mathbf{C} - \lambda \mathbf{I}) = 0 \tag{5.1.1.6}$$

where $\mathbf{I}$ is the identity matrix, and solving this provides the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. Once the eigenvalues are known, the corresponding eigenvectors $\mathbf{v}_i$ can be computed by solving the linear system $(\mathbf{C} - \lambda_i \mathbf{I})\mathbf{v}_i = 0$ for each λ_i.

4. **Principal Components Selection**: The components are then ranked with the eigenvalues in descending order and the top k eigenvectors are selected and explored further. The choice of k depends on the desired level of variance to be retained, where more variance retained requires a larger number of components to be selected.

5. **Transformation**: Using the top k selected components, the original data is then transformed into the new subspace. This results in a reduced set of variables (principal components) that capture the most significant features of the data.

$$\mathbf{S} = \mathbf{X}\mathbf{W} \tag{5.1.1.7}$$

where $\mathbf{W}$ is the $p \times k$ matrix of the top k eigenvectors, and $\mathbf{S}$ is a $n \times k$ reduced subspace of new latent features.

6. **Interpretation and Analysis**: The $p \times k$ dataset, S can then be analysed to glean insights, visualize patterns, and build models based on a dataset with reduced dimensions.

5.1.1.4 Implementation

The Python code needed to implement PCA using the `scikit-learn` (Pedregosa et al., 2011) library can be found on the Github repository for this project: https://github.com/WesleyJKing/thesis.git.

5.1.2 T-SNE

This section is based off the paper "*Visualizing Data using t-SNE*" by van der Maaten and Hinton (2008).

5.1.2.1 Introduction

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a state-of-the-art dimension reduction technique which focuses on reducing the dimensions of the data, while maintaining the relative distance between observations throughout the dimension reduction process. t-SNE is an improvement on the work done by Hinton and Roweis (2002) who created Stochastic Neighbor Embedding (SNE). SNE is hampered by a cost function (measures the divergence between high-dimensional and low-dimensional probability distributions) that is difficult to optimize, along with the Crowding Problem (refer to Hinton and Roweis (2002) for more information on the crowding problem). To combat this, t-SNE utilizes a symmetrized version of the cost function, and a heavy-tailed Student t-distribution rather than a Gaussian, to compute similarities of data points in the lower dimension. The perplexity parameter within the cost function is the only parameter that necessitates tuning. Conceptually, perplexity can be regarded as the effective number of neighbors. Perplexity, typically set within the range of 5 to 50, plays a crucial role in determining the desired scope of analysis. Thus, when perplexity is set to a small value closer to 5, the emphasis is placed on capturing local structure by considering fewer neighboring points. Conversely, a larger perplexity value shifts the focus towards capturing the global structure of the data.

5.1.2.2 Underlying Assumptions

t-SNE operates under the following key assumptions:

- **Local Structure Preservation**: The method assumes that the local structure of the data is more important than the global structure. Generally, it emphasizes preserving the distances between nearby data points in the reduced-dimensional space. However, the extent of this can be controlled with the adjustment of the perplexity parameter.

- **Non-linear Relationships**: t-SNE is designed to capture non-linear relationships in the data, making it effective for visualizing complex, high-dimensional data.
- **Gaussian Distribution in High Dimensions**: The pairwise distances between points in the high-dimensional space are modeled using a Gaussian distribution.
- **Student's t-Distribution in Low Dimensions**: The pairwise distances in the low-dimensional space are modeled using a Student's t-distribution with one degree of freedom.

5.1.2.3 Algorithm and Mechanism

The t-SNE algorithm consists of the following steps:

1. **Compute Pairwise Affinities in High Dimensions**: t-SNE uses a Gaussian kernel to calculate the pairwise affinities (similarities) between data points in the high-dimensional space where the affinity p_{ij} between points x_i and x_j is given by:

$$p_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{5.1.2.1}$$

   where σ_i is the bandwidth of the Gaussian kernel for point x_i. In this context, bandwidth σ_i controls the spread of the Gaussian distribution centered at x_i, effectively determining how local or global the similarity measure is.

2. **Symmetrize Affinities**: The affinities are symmetrized to ensure that $p_{ij} = p_{ji}$. The symmetrized affinity P_{ij} is given by:

$$P_{ij} = \frac{p_{ij} + p_{ji}}{2N} \tag{5.1.2.2}$$

   where N is the total number of observations.

3. **Compute Pairwise Affinities in Low Dimensions**: The points in the low-dimensional space are initialized and the pairwise affinities (q_{ij}) are computed using a Student's t-distribution with one degree of freedom:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{5.1.2.3}$$

   where y_i and y_j are the low-dimensional representations of x_i and x_j.

4. **Minimize Kullback-Leibler Divergence**: Minimize the Kullback-Leibler (KL) divergence between the high-dimensional affinities P_{ij} and the low-dimensional affinities Q_{ij}. The KL divergence is given by:

$$KL(P \parallel Q) = \sum_{i \neq j} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{5.1.2.4}$$

This minimization is done using gradient descent, where the gradient of the KL divergence with respect to the low-dimensional points y_i is given by:

$$\frac{\partial KL}{\partial y_i} = 4 \sum_j (P_{ij} - Q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \tag{5.1.2.5}$$

Update the positions of the points in the low-dimensional space to reduce the KL divergence using the gradient descent:

$$y_i^{(t+1)} = y_i^{(t)} - \eta \frac{\partial KL}{\partial y_i} + \alpha(t)(y^{(t)} - y^{(t-1)}) \tag{5.1.2.6}$$

where η is the learning rate(regulates the magnitude of the updates, balancing convergence speed and stability) and α(t) is a momentum term (stabilizes and accelerates optimization by incorporating previous updates) at iteration t.

5.1.2.4 Implementation

The Python code needed to implement t-SNE using the `scikit-learn` library can be found on the Github repository for this project: https://github.com/WesleyJKing/-thesis.git.

5.1.3 Autoencoders

This section is based on the foundational work of "*Reducing the Dimensionality of Data with Neural Networks*" by Hinton and Salakhutdinov (2006).

5.1.3.1 Introduction

Autoencoders are a subset of artificial neural networks used to train latent representations of data in an unsupervised manner. The objective of an autoencoder is to transform the inputs into a reduced representation and then reconstruct the original input from this representation. This process forces autoencoders to identify important features and patterns in the data, making them valuable for dimensionality reduction, feature learning, and anomaly detection. Figure 28 below shows that autoencoders consist of two main parts: an encoder that compresses the input

(x_1, x_2, x_3) into a latent-space representation z, and a decoder that reconstructs the input from this latent representation to form the output $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$. The learning process aims to minimize the difference between the original input and its reconstruction.

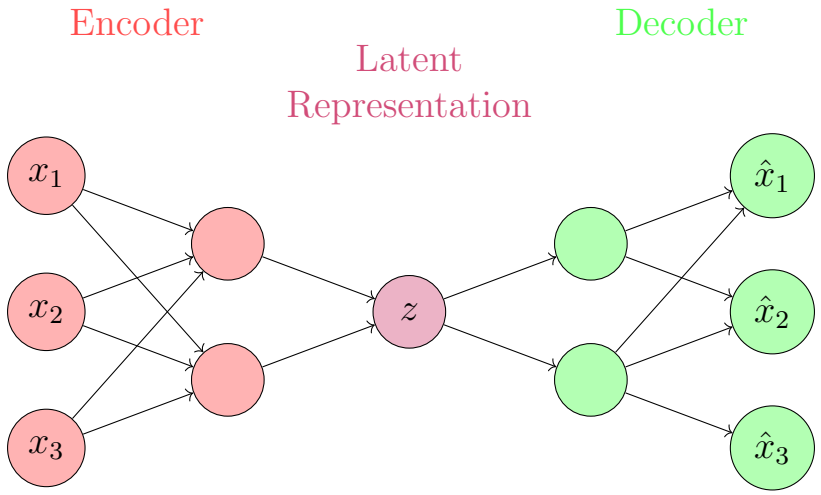

Figure 28: A basic representation of an Autoencoder neural network

5.1.3.2 Underlying Assumptions

Autoencoders operate under the following key assumptions:

- **Data Compression**: The model assumes that the input data can be compressed into a lower-dimensional representation without significant loss of important information.
- **Reconstruction Capability**: It is assumed that the autoencoder can reconstruct the input data accurately from its compressed representation, capturing the underlying structure and features of the data.
- **Non-linear Transformations**: Autoencoders leverage non-linear transformations, making them capable of capturing complex patterns and relationships within the data.
- **Symmetry between Encoder and Decoder**: The encoder and decoder are typically symmetric, meaning they have similar architectures but reversed roles in terms of data transformation.

5.1.3.3 Algorithm and Mechanism

The autoencoder algorithm consists of the following steps:

1. **Input Layer**: The input data $\mathbf{x}$ is fed into the encoder network. Each input vector $\mathbf{x} \in R^n$ is transformed into a hidden representation $\mathbf{h} \in R^m$ (where $m < n$) through the encoder

$$\mathbf{h} = f(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e) \tag{5.1.3.1}$$

   where f is a non-linear activation function, $\mathbf{W}_e$ are the weights of the encoder, and $\mathbf{b}_e$ is the bias.

2. **Latent Space Representation**: The hidden representation $\mathbf{h}$ captures the essential features of the input data in a lower-dimensional space.

3. **Decoder Network**: The decoder transforms the hidden representation $\mathbf{h}$ back into a reconstruction $\hat{\mathbf{x}}$ of the input data

$$\hat{\mathbf{x}} = g(\mathbf{W}_d\mathbf{h} + \mathbf{b}_d) \tag{5.1.3.2}$$

   where g is a non-linear activation function, $\mathbf{W}_d$ are the weights of the decoder, and $\mathbf{b}_d$ is the bias.

4. **Loss Function**: The autoencoder minimizes the reconstruction error between the input data $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. The most commonly used loss function is the mean squared error (MSE)

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2. \tag{5.1.3.3}$$

5. **Training Process**: The weights of the encoder and decoder are optimized using gradient descent or other optimization algorithms to minimize the reconstruction error

$$\mathbf{W}_e, \mathbf{W}_d, \mathbf{b}_e, \mathbf{b}_d \leftarrow \text{argmin}_{\mathbf{W}_e, \mathbf{W}_d, \mathbf{b}_e, \mathbf{b}_d}\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) \tag{5.1.3.4}$$

5.1.3.4 Implementation

The Python code needed to implement this using the `pytorch` (Paszke et al., 2019) library can be found on the Github repository for this project: https://github.com/-WesleyJKing/thesis.git.

5.2 Predictive Models

5.2.1 Random Forest

This section is based off the "*Random Forests*" paper by Breiman (2001b).

5.2.1.1 Introduction

The Random Forest algorithm is an ensemble learning method, which combines multiple weak learners, typically decision trees, to improve overall predictive performance by reducing variance and increasing robustness and is used for classification and regression tasks. Random Forests consist of trees where each tree is a hierarchical structure composed of internal nodes and leaves. Internal nodes contain decision rules that split the data based on feature thresholds, while leaves are the terminal nodes that contain the actual predictions. As data flows through the tree, it passes through these internal nodes, following the appropriate split paths until reaching a leaf node. These trees are optimized during training, and each outputs its own prediction, with the final output determined by averaging (for regression) or majority voting (for classification). Random Forests improve predictive accuracy and mitigate overfitting problems when compared to individual decision trees by combining many diverse trees, each seeing different subsets of the data and features.

5.2.1.2 Underlying Assumptions

Random Forests operate under the following key assumptions:

- **Feature Independence:** Variables are assumed to be independent, as the primary idea of random forests is to minimize the correlation between decision trees without substantially increasing the variance.
- **Generalization:** Generalizability is achieved by constructing multiple uncorrelated decision trees, where each tree is trained on a randomly sampled subset of the data and only a subset of features is considered at each split. This approach reduces overfitting and improves model stability by averaging predictions across diverse trees. This technique is referred to as *bagging*.

5.2.1.3 Algorithm and Mechanism

The Random Forest algorithm involves the following steps:

1. **Bootstrap Sampling**: Bootstrapping is a resampling technique that creates multiple datasets by randomly selecting data points from the original dataset with replacement. This allows each bootstrap sample to contain duplicate instances while maintaining the same overall size as the original dataset. Multiple bootstrap samples are then generated, each serving as a training set for a different model instance. The final bootstrap dataset looks as follows:

$$D_b = \{(x_i, y_i)\}_{i=1}^{N_b} \tag{5.2.1.1}$$

where D_b is the b-th bootstrap sample to be used to create the b-th decision tree, and N_b is the number of data points in the b-th sample.

2. **Grow Decision Trees**: For each bootstrap sample, grow a decision tree. During the construction of the decision trees, at each node, select a random subset of features and determine the best variable to split on based on the chosen impurity measure, which evaluates how well a split separates the data by their target values:

$$\text{Variable} = \arg\min_{f \in F_{sb}} \text{Impurity}(D_b) \tag{5.2.1.2}$$

where F_{sb} is the random subset of features for the s-th split of the b-th tree, and Impurity(D_b) is the chosen impurity measure (e.g., Gini impurity, entropy for classification or MSE, MAE for regression tasks). This dissertation used MSE, which means splits were evaluated based on how much they reduced the average squared difference between the predicted values and the actual target values in each resulting node. A lower MSE after splitting indicates that the data points within each node are more similar in terms of their target values, leading to a better division of the data.

3. **Repeat**: Repeat the bootstrap sampling and tree growing process to create B decision trees

$$\text{Forest} = \{T_1, T_2, \ldots, T_B\} \tag{5.2.1.3}$$

where T_i is the i-th decision tree.

4. **Generate Out-of-Bag (OOB) Predictions**: For each observation x_i for $i = 1, 2, ..., N$, aggregate the predictions across all trees where this observation wasn't included in the bootstrap sample for that tree. For regression, aggregate the OOB predictions by taking the mean, and for classification, aggregate the OOB predictions by taking a majority vote:

$$\hat{y}_{\text{OOB}} = \frac{1}{|B_{\text{OOB}}(x_i)|} \sum_{b \in B_{\text{OOB}}(x_i)} T_b(x_i) \quad \text{(regression)} \tag{5.2.1.4}$$

and

$$\hat{y}_{\text{OOB}} = \text{mode}\{T_b(x_i) \mid b \in B_{\text{OOB}}(x_i)\} \quad \text{(classification)} \tag{5.2.1.5}$$

where $T_b(x_i)$ is the prediction of the b-th tree for input x_i, and $B_{\text{OOB}}(x_i)$ is the set of trees for which x_i was out-of-bag.

5.2.1.4 Implementation

The Python code needed to implement this using the `scikit-learn` library can be found on the Github repository for this project: https://github.com/WesleyJKing/thesis-

.git.

5.2.2 OLS Regression

This section is grounded in the foundational principles of linear regression as notably described in "*Statistical Models: Theory and Practice*" by Freedman (2009).

5.2.2.1 Introduction

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between a dependent variable and multiple independent variables. The origins of linear regression trace back to the late 19th century with Sir Francis Galton, a polymath who sought to understand heredity through his experiments with sweet peas. By plotting the sizes of offspring seeds against parent seeds, Galton discovered the concept of "regression to the mean," where extreme traits in one generation tend to average out in the next. This discovery laid the groundwork for modern regression analysis.

Karl Pearson, a contemporary of Galton and a significant figure in the development of statistics, expanded on Galton's ideas by providing a mathematical foundation for correlation and regression. Pearson's work formalized the methods that Galton had intuitively developed, enabling the application of regression analysis to a broader range of problems beyond heredity, including psychology, economics, and social sciences. Pearson's rigorous treatment of these concepts ultimately led to the development of MLR as a powerful tool for analyzing linear relationships between variables.

Initially applied to biological data, MLR has since become one of the most widely used statistical methods across various fields. It allows researchers to quantify the impact of multiple factors on an outcome, facilitating its widespread use in disciplines such as finance, medicine, and social sciences.

5.2.2.2 Underlying Assumptions

The validity of the Multiple Linear Regression model relies on several key assumptions:

- **Linearity**: The relationship between the dependent variable and each of the independent variables is linear.
- **Multicollinearity**: Variables are independent of each other.
- **Homoscedasticity**: The variance of the errors is assumed to be constant across all values of the independent variables.
- **Normality of Errors**: The errors are assumed to be normally distributed.

5.2.2.3 Algorithm and Mechanism

The procedure for fitting a Multiple Linear Regression model typically involves the following steps:

1. **Model Specification**: The model can be defined as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{5.2.2.1}$$

where y is the dependent variable, $x_1, x_2, \ldots, x_p$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients, and ϵ represents the error term. This can also be written in matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.2.2.2}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

2. **Parameter Estimation**: The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by minimizing the sum of squared errors (SSE) between the observed y and the predicted values $\hat{y}$ of the dependent variable. This can be achieved through Ordinary Least Squares (OLS):

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{5.2.2.3}$$

where X is the matrix of independent variables, and y is the vector of observed values of the dependent variable.

3. **Model Evaluation**: The Mean Squared Error (MSE) quantifies the average of the squared differences between the observed and predicted values, providing a measure of the model's prediction accuracy:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5.2.2.4}$$

where $\hat{y}_i$ are the predicted values, y_i are the observed values, and $\bar{y}$ is the mean of the observed values. Thus, a lower MSE indicates a better fit of the model to the data, as it reflects smaller discrepancies between the predicted and observed values.

Additionally, another critical element needed to assess the fit of the linear model is the coefficient of determination, denoted as R^2. This is a measure used to assess how well a regression model explains the variance in the dependent variable. It is thus defined as the proportion of the variance in the dependent variable that is predictable from the set of independent variables, where R^2 values range from 0 to 1, where, where higher values of R^2 indicate a better fit. The formula for R^2 can be seen as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.2.2.5}$$

where y_i is the i-th value of the dependent variable, $\hat{y}_i$ is the predicted i-th value, $\bar{y}$ is the mean of the actual values, and n is once again the number of observations. The numerator is thus the Residual Sum Of Squares (RSS) and the denominator is a measure of the total variability in the dependent variable. One downfall of this approach is that any added variables will likely increase R^2, and definitely not decrease it. Thus, there are no penalties for having unnecessary variables. For this reason, R^2_{adj} will be analysed:

$$R^2_{\text{adj}} = 1 - \left(\frac{(1 - R^2)(n-1)}{n-p-1}\right). \tag{5.2.2.6}$$

This formula adjusts R^2 by accounting for the number of variables (p) and the sample size N, penalizing the inclusion of unnecessary variables. Unlike R^2, Adjusted R^2 can decrease if the added variables do not sufficiently enhance the model's explanatory power relative to the complexity they introduce (Chen & Qi, 2023). This feature promotes the selection of simpler, more parsimonious models, aligning with Occam's Razor in statistical modeling (Bargagli Stoffi et al., 2022).

Moreover, during variable selection, Akaike's Information Criterion (Akaike, 1974) will be used:

$$\text{AIC} = 2k - 2\ln(L), \tag{5.2.2.7}$$

where k is the number of parameters in the model, while L is the natural logarithm of the likelihood function. This means the number of variables included in the model increases, $2k$ will increase, as well as $2\ln(L)$, since new variables will likely increase the likelihood function assuming the variables hold some explanatory power. Conversely, if variables are removed from the model the first term will decrease, but the likelihood function will also decrease. In this way, AIC gives an intuitive trade-off between model complexity and explanatory power, encouraging a parsimonious model that also explains the data well. It is important to note that AIC is used over The Bayesian information criterion (BIC) because BIC is primarily used when one assumes the true model is in the scope of all models that are possible within the current dataset (Schwarz, 1978).

Another evaluation metric that will be utilised is Precision@k which can be defined as the proportion of relevant items among the top k items that are indeed in the top k (Hicks et al., 2022), mathematically expressed as:

$$\text{Precision@}k = \frac{\text{Relevant Items in Top-}k}{k}. \tag{5.2.2.8}$$

Precision@k will provide a more heuristic measure of how well the models rank players relative to their actual OBV ranks. This essentially simulates a scout's behaviour and quantifies whether the top players are indeed found amongst the predicted top k players, and what percentage of them are actually correct. Since there are many leagues to scout from, and a primary aim of this paper is to encourage a concise list of potential recruits, k will be analysed for small values of 5 and 10.

4. **Assumption Checking**: Verify the underlying assumptions of linearity, independence, homoscedasticity, and normality of errors through diagnostic plots and statistical tests.

5. **Model Refinement**: If the model assumptions are violated, consider transformations of the independent variables, adding interaction terms, or using regularization techniques like Ridge or Lasso Regression to improve model performance and robustness.

6. **Prediction**: Use the fitted model to make predictions on new data, by plugging in the values of the independent variables into the estimated regression equation.

Multiple Linear Regression provides a straightforward yet powerful tool for understanding and predicting the relationship between a dependent variable and multiple independent variables. It remains one of the most commonly used methods in both academic research and industry applications.

5.2.2.4 R Implementation

The R implementation was performed using the `lm()` function in the `stats` package (R Core Team, 2022).

5.2.3 Multi-Layer Perceptron (MLP)

This section is based on the foundational work presented in "*Learning representations by back-propagating errors*" by Rumelhart et al. (1986).

5.2.3.1 Introduction

An MLP is a computational model inspired by the way biological neural networks in the human brain process information, an example of which is shown in Figure 29 below. It consists of interconnected layers of nodes, or neurons, where each layer transforms the input data (x_1, x_2, x_3) using weighted connections and activation functions to generate the output. An activation function applies a mathematical transformation to the output of a neuron. By introducing non-linearity, it allows the network to learn and model complex, non-linear relationships in the data. Additionally, the bias terms, represented by nodes with a value of 1 in the network, are multiplied by an optimized weight and added to the weighted input. This ensures that the bias term provides flexibility to shift the activation function appropriately, allowing the model to better fit the data even when all inputs are zero.

5.2.3.2 Underlying Assumptions

Neural networks operate under the following key assumptions:

- **Continuity and Differentiability:** The activation functions used within the neurons are typically continuous and differentiable, allowing the use of gradient-based optimization methods.
- **Large Data Requirement:** Neural networks assume the availability of large datasets for effective training, as they have a high capacity to learn complex patterns.
- **Data Representation:** Input data is often assumed to be normalized or standardized, and features are assumed to be relevant and informative for the

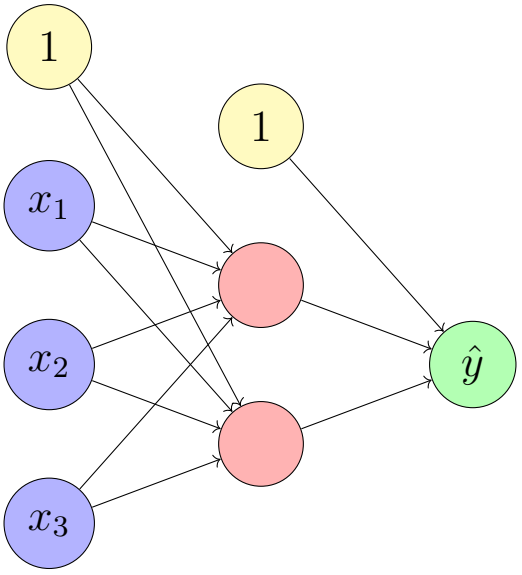

Figure 29: A neural network diagram showing input nodes, a hidden layer with bias, and the output node.

task at hand.

5.2.3.3 Algorithm and Mechanism

The neural network algorithm involves the following steps:

1. **Initialization**: Initialize the weights and biases of the network, typically with small random values. For the case of this example, we can let the initial weights and biases be denoted as:

$$W_{jk}^{(l)} \sim \mathcal{N}(0, \sigma^2), \quad b_j^{(l)} \sim \mathcal{N}(0, \sigma^2) \tag{5.2.3.1}$$

   where $W_{jk}^{(l)}$ and $b_j^{(l)}$ are the weights and biases of the l-th layer, where the weight connects the k-th neuron in the l-th layer to the j-th neuron in layer $l+1$.

2. **Forward Propagation**: For each input x_i, compute the output of each layer using the activation function. The output of the l-th layer is given by:

$$z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l, \tag{5.2.3.2}$$

   and

$$a_j^l = \sigma^l(z_j^l), \tag{5.2.3.3}$$

where z_j^l is the pre-activated output of the j-th neuron in the l-th layer, while σ^l is the activation function used in the l-th layer (e.g., sigmoid, ReLU). Some of the most widely used activation functions include:

- **Sigmoid**: The sigmoid activation function maps inputs to the range $(0, 1)$, making it useful for binary classification

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{5.2.3.4}$$

- **Tanh**: The hyperbolic tangent function maps inputs to $(-1, 1)$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{5.2.3.5}$$

- **ReLU (Rectified Linear Unit)**: A widely used activation function in hidden layers, it simplifies training by setting negative inputs to zero

$$\text{ReLU}(x) = \max(0, x). \tag{5.2.3.6}$$

- **Softmax**: Commonly used in multi-class classification, the softmax function converts logits into probabilities

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}. \tag{5.2.3.7}$$

3. **Loss Calculation**: Compute the Loss Function which quantifies the difference between the predicted value and the actual target. Some common loss functions include:

- **Mean Squared Error (MSE)**: Used for regression tasks, MSE penalizes large errors more heavily

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5.2.3.8}$$

where y_i is the actual target, $\hat{y}_i$ is the predicted output, and N is the number of observations.

- **Binary Cross-Entropy (BCE)**: Used for binary classification tasks, BCE measures the logarithmic loss between predicted and actual probabilities

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right). \tag{5.2.3.9}$$

- **Categorical Cross-Entropy (CCE)**: Used for multi-class classification tasks, CCE generalizes BCE by summing over all classes

$$\text{CCE} = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log(\hat{y}_{i,k}) \tag{5.2.3.10}$$

where k is an index for the class. For a regression task, the mean squared error (MSE) with L2 regularization (also known as Ridge Regression) can be used:

$$\mathcal{C} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{P} w_j^2 \tag{5.2.3.11}$$

where λ is the regularization parameter, w_j are the weights of the model, and P is the number of weights. L2 regularization in the loss function penalizes weights with large magnitudes, effectively shrinking them. This parameter needs to be optimized to enhance the model's generalizability, thereby improving performance while preventing overfitting.

4. **Backward Propagation**: Backward Propagation involves the ultimate goal of computing the gradient of the loss function with respect to each weight and bias using the chain rule. In order to do this however, the following equations must first be understood

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \tag{5.2.3.12}$$

and

$$\nabla_a C = \begin{pmatrix} \frac{\partial C}{\partial a_1^L} \\ \frac{\partial C}{\partial a_2^L} \\ \vdots \\ \frac{\partial C}{\partial a_m^L} \end{pmatrix}. \tag{5.2.3.13}$$

This equation calculates the error term δ^L at the output layer (layer L). The error is the element-wise product (denoted by $\odot$) of the gradient of the cost function with respect to the activation, $\nabla_a C$, and the derivative of the activation function, $\sigma'(z^L)$. This step is crucial for propagating the error backward from the output layer

$$\delta^l = \left((w^{l+1})^T\delta^{l+1}\right) \odot \sigma'(z^l). \tag{5.2.3.14}$$

This equation propagates the error backwards from layer $l+1$ to layer l. It computes the error term δ^l for layer l as the element-wise product of the backpropagated error $(w^{l+1})^T\delta^{l+1}$ and the derivative of the activation function at layer l, $\sigma'(z^l)$. Here, $(w^{l+1})^T$ is the transpose of the weight matrix for layer $l+1$

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l. \tag{5.2.3.15}$$

This equation expresses that the gradient of the cost function with respect to the bias b_j^l at layer l is simply the error term δ_j^l. This is because the bias influences the cost function directly through the error term

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l. \tag{5.2.3.16}$$

This equation shows that the gradient of the cost function with respect to the weight w_{jk}^l at layer l is the product of the activation a_k^{l-1} from the previous layer (layer $l-1$) and the error term δ_j^l at the current layer. This reflects how much each weight contributes to the error in the output.

5. **Gradient descent**: Update the weights and biases using gradient descent:

$$w_{ij}^l \leftarrow w_{ij}^l - \eta\frac{\partial \mathcal{C}}{\partial w_{ij}^l}, \quad b_i^l \leftarrow b_i^l - \eta\frac{\partial \mathcal{C}}{\partial b_i^l} \tag{5.2.3.17}$$

where η is the learning rate. The learning rate controls the size of the steps taken during the optimization process to minimize the loss function. A smaller learning rate ensures gradual, precise updates, but may slow down convergence, while a larger learning rate speeds up convergence but risks overshooting the optimal solution or causing instability.

6. **Iteration and Convergence**: Repeat the forward and backward propagation steps for a specified number of epochs or until the loss function converges to a minimum value.

7. **Prediction**: Once the network is trained, use it to make predictions on new data by performing forward propagation with the learned weights and biases.

5.2.3.4 Implementation

The Python code needed to implement this using the `pytorch` (Paszke et al., 2019) library can be found on the Github repository for this project: https://github.com-/WesleyJKing/thesis.git.

5.2.4 Extreme Gradient Boosted Trees (XGBoost)

This section is based on the paper "*XGBOOST: A Scalable Tree Boosting System*" by Chen and Guestrin (2016a).

5.2.4.1 Introduction

XGBoost (Extreme Gradient Boosting) represents a highly efficient and scalable implementation of gradient boosting algorithms. It is a powerful machine learning technique that builds a sequence of decision trees, where each new tree learns from and improves upon the mistakes of its predecessors. At its core, it uses gradient-based boosting to enhance accuracy, while incorporating sophisticated features like regularization and parallel processing. XGBoost is not merely an enhancement over traditional boosted trees; it introduces innovative algorithmic and system optimizations, making it a top choice for tasks in both classification and regression domains.

5.2.4.2 Underlying Assumptions

XGBoost operates under several key assumptions:

- **Additive Models**: The model construction follows an additive framework where trees are sequentially added to the model to reduce prediction errors incrementally.
- **Differentiable Loss Function**: The loss function employed in XGBoost is assumed to be differentiable. This characteristic is crucial for the application of gradient-based optimization techniques, such as gradient descent, which are necessary for refining the model's parameters.

5.2.4.3 Algorithm and Mechanism

The XGBoost algorithm is structured around several core components that contribute to its efficiency and effectiveness, particularly in handling cases where many variables have values of zero (sparsity) and optimizing memory usage. Below is an overview of the primary aspects of the algorithm:

1. **Regularized Learning Objective**: For a dataset comprising n examples and m features $D = \{(x_i, y_i)\}$, where $|D| = n$, $x_i \in R^m$, and $y_i \in R$, the tree ensemble model in XGBoost uses K additive functions to predict the output:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{5.2.4.1}$$

Here, $\hat{y}_i$ represents the predicted value for the i-th data point, while x_i is its corresponding feature vector, consisting of m input features. The function space $\mathcal{F} = \{f(x) = w_{q(x)}\}$, with $q : R^m \to T$ mapping input features to one of T leaf nodes, and $w \in R^T$ representing the corresponding leaf weights, defines the space of regression trees. Here, T denotes the number of leaf nodes in each regression tree, meaning that each function $f_k(x_i)$ represents a tree in the ensemble, where the prediction is determined by the weight associated with the leaf node assigned by $q(x_i)$.

2. **Objective Function**: The objective function in XGBoost can be defined as:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5.2.4.2}$$

where $l(y_i, \hat{y}_i)$ represents the loss function which measures the difference between the predicted value and the actual value, and $\Omega(f_k)$ is the regularization term which penalizes complex models (many leaf nodes), promoting more parsimonious ones.

3. **Gradient and Hessian Calculation**: To optimize the objective, XGBoost utilizes a second-order Taylor expansion of the loss function and computes the first and second derivatives (gradient and Hessian) with respect to the predictions from the previous boosting iteration:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}} \tag{5.2.4.3}$$

where g_i and h_i are the first and second-order gradient statistics of the loss function. These derivatives allow XGBoost to approximate the loss function using a second-order Taylor expansion, leading to efficient tree optimization and faster convergence.

4. **Tree Structure Score**: To optimize the tree, XGBoost assigns weights to each leaf node and calculates a score function to determine the best structure. Given a fixed tree structure q and a set of leaf weights w_j, the optimal weight for each leaf j is derived from the second-order Taylor expansion of the loss function:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{5.2.4.4}$$

where I_j represents the set of data points assigned to leaf j, g_i and h_i are the first and second-order gradients of the loss function, and λ is an L2 regularization parameter that prevents overfitting.

Using this optimal weight, the quality of a given tree structure is measured using the tree structure score:

$$\text{Score}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{5.2.4.5}$$

where T is the number of leaf nodes in the tree, and γ is a penalty parameter that discourages over-complex trees by penalizing large T. The score function, derived from a second-order approximation of the loss function, plays a crucial role in split evaluation and pruning. XGBoost uses this score to assess potential splits at each step of tree construction. A higher score indicates a better split that contributes to reducing the overall loss, while the regularization term γT ensures that unnecessary splits are avoided. By comparing the score before and after a split, XGBoost determines whether the split improves the tree sufficiently or should be pruned.

5.2.4.4 Implementation

The Python code needed to implement this using the `XGBOOST` library (Chen & Guestrin, 2016b) can be found on the Github repository for this project: https://-github.com/WesleyJKing/thesis.git.

5.3 Data Preparation

The initial stage of data processing involved retrieving data from multiple sources. This began with extracting all available *OBV* data using the Statsbomb API (Hudl StatsBomb, 2021b) . The seasons for which *OBV* data was available were recorded, and the relevant explanatory variables for the players during these seasons were manually downloaded from the Wyscout (Wyscout, 2024) and FBref (FBref, 2025) websites.

Informed by the Exploratory Data Analysis, the steps taken to transform the data were:

1. **One-Hot Encoding:** The sole categorical variable, *Position.category*, was one-hot encoded to convert it into a wide numerical format. The original *Position.category* column was then dropped from the dataset.
2. **Feature Scaling:** The numerical features in the dataset (i.e., all variables excluding the previously created position-encoded variable) were scaled, ensuring

that each feature had a mean of 0 and a standard deviation of 1. Scaling numerical variables is important because it improves algorithm efficiency, speeds up convergence in optimization processes like gradient descent, and enhances model accuracy by preventing bias towards those with larger ranges (Ahsan et al., 2021).

3. **Combining Features:** The scaled numerical features were then concatenated with the one-hot encoded categorical features.

4. **Data Splitting:** The dataset was then split into three subsets: training, validation, and test sets, using a 70/15/15 split. This was done to allow for model training, validation, and evaluation on distinct portions of the data.

5.3.1 Variable Selection Methods for ML Models

To optimize model performance and reduce the dimensionality of the dataset, two approaches were employed: Principal Component Analysis (PCA) and feature importance-based selection. PCA was applied to reduce the feature space by transforming the original variables into a set of orthogonal components that capture the maximum variance in the data. This approach was selected due to the presence of cases where variables were correlated with each other, as was found in the EDA. Two PCA informed datasets were formed: the first retained the minimum number of components that explained over 99% of the total variance, while the second kept the minimum number of components that cumulatively explained at least 80% of the variance. These thresholds were chosen to balance the trade-off between preserving essential variance and significantly reducing the number of input features.

Feature selection was performed using variable importance metrics derived from both XGBoost and Random Forest models. In this process, each feature's contribution to the model's predictive power was evaluated, and variables were selected according to three different strategies:

1. **Threshold-based selection:** Features were selected if their importance score exceeded a pre-specified threshold of 0.05 (Wang et al., 2024), ensuring that irrelevant variables were removed.

2. **Top-N selection:** This method retained the top N features based on their importance ranking, where N is a chosen number of the most significant features. The value of N depended on the feature set size, but the proportion selected was 80%.

3. **Cumulative importance selection:** Features were selected until the cumulative importance exceeded a defined threshold of 95%. This method ensured

that the most important variables were captured while discarding those with minimal contribution to the model's output.

The feature selection process was iterative and involved numerous methods to allow for many combinations of variables to be tested to provide the model with an optimal environment for high performance. This helped to streamline the model, improving both computational efficiency and predictive accuracy.

5.4 Model building

5.4.1 OLS Regression

In the development of the linear models, a top-down approach was implemented since hypotheses regarding the relationships between explanatory variables and the outcome variable had previously been establised in the EDA (Ley et al., 2022). Thus, the objective was to simplify the model by systematically removing less significant predictor variables while improving model fit, as primarily measured by the Akaike Information Criterion (AIC), but also considering adjusted R^2. The process began by creating a full linear model, including all variables if OBV is being modelled in its entirety, otherwise including all variables associated with the relevant OBV component. This overall model is then assessed by removing variables that contribute little to the predictive power based on a step-wise AIC approach. Specifically, each iteration sees the evaluation of whether removing a specific variable lowers the AIC. This is evaluated for each variable in the current model. If the AIC is lowered after the removal of a variable, this indicates an improvement in the model's balance between complexity and fit. In each iteration, the variable which reduces the AIC most is removed. Through repeated iterations, the model is progressively refined by eliminating the least significant variables according to the AIC. This continues until further removals fail to improve the AIC, at which point the model is considered optimal. The result is a more parsimonious model that retains only the most significant predictors, leading to better generalizability and interpretability.

5.4.2 Multi-Layer Perceptron

Modeling with MLPs was a complex task that required an iterative process of fine-tuning and refining the model. Thus, to determine the optimal neural network architecture for modeling the relevant OBV component, as well as the *Total OBV*, a comprehensive process was undertaken involving the generation, training, validation, and evaluation of numerous neural network configurations.

Firstly, numerous neural networks were developed to systematically create a range of neural network architectures of differing sizes and activation functions. To do

this effectively, custom functions were created which built neural network architectures with permutations of hyperparameters. These parameters included: L2-regularization, activation functions, and the ratio parameter, which is a custom variable correlated with the number of layers in the network. Descriptions for each optimized hyperparameter are shown below:

1. **Lambda (L2-regularization):** L2-regularization involves adding a penalty proportional to the square of the model coefficients to the loss function, encouraging smaller weights and reducing model complexity. This helps prevent overfitting by balancing the bias-variance tradeoff, which reflects the fundamental tradeoff between model complexity and generalization. Higher regularization reduces variance by simplifying the model, making it less sensitive to fluctuations in the training data but increasing bias by limiting its ability to capture complex patterns. Conversely, lower regularization reduces bias by allowing the network to learn more intricate relationships, but at the cost of increased variance, making it more susceptible to overfitting. To find the optimal Lambda value, the model validation loss function is plotted on the y-axis, and the lambda values are on the x-axis. It is important that the loss function be somewhat convex. Generally, such a plot shows large loss values for under-regularized models (suggesting overfitting), small loss values for somewhat regularized models, and then increasing loss values as the lambda values climb further, suggesting highly constrained models.
2. **Activation Functions:** The activation functions experimented with included the identity function, the Sigmoid function, LeakyReLU function, ReLU function, and the Tanh function. This allowed for possible linear transformations to be modeled, as well as different non-linear transformations.
3. **Ratio:** The ratio parameter describes how fast the data reduction occurs to the output layer. Thus, a value of 0.9 means layer $n + 1$ will have 0.9 times the number of neurons in layer n, or:

$$n_{(n+1)} = 0.9 \times n_n \tag{5.4.2.1}$$

which results in a neural network with a large number of layers. Conversely, a ratio value of 0.1 means layer $n + 1$ will have 10% of the neurons that the previous layer had, or:

$$n_{(n+1)} = 0.1 \times n_n \tag{5.4.2.2}$$

To find the optimal number of layers, the parameter grid search tested values

of the ratio parameter between 0.1 and 0.9.

To select the optimal model, we first split the data into training, validation, and test sets in a 70-15-15 split. A series of training and validation loops was performed on each MLP, using a range of L2 regularization values to prevent overfitting, before they were tested on the test set. Each neural network was trained over a specified number of epochs using the Adam optimizer and mean squared error (MSE) loss function. During training, the model's performance was periodically evaluated on the validation set to monitor its progress and avoid overfitting. Training, and validation losses were recorded for each epoch, and the configuration yielding the lowest validation loss was identified as the optimal model.

5.4.3 Random Forest

Random Forests require extensive tuning, as they have numerous parameters that need to be optimized to achieve peak performance. A 70-15-15 data split was used to train, validate, and test the Random Forest models. A parameter grid was defined to comprehensively search the parameter space, specifying ranges for the following parameters:

1. **Number of Estimators:** Determines the number of trees in the forest. Increasing the number of trees can enhance the model's accuracy by averaging out the predictions of individual trees, which reduces variance. However, beyond a certain point, adding more trees results in diminishing returns, where the improvement in accuracy becomes negligible while computational costs increase (Probst & Boulesteix, 2017). To cover different magnitudes of estimators, 100 and 120 estimators were tested.

2. **Maximum Depth:** Controls the maximum depth of the trees. A deeper tree allows the model to capture more complex patterns but increases the risk of overfitting. Conversely, a shallower tree might underfit by failing to capture sufficient detail in the data. Maximum depths of 8, 10, and 12 were tested.

3. **Minimum Samples Split:** Specifies the minimum number of samples required to split an internal node. A higher value makes the model more conservative, reducing the likelihood of splits that might overfit to noise in the data. A lower value allows splits to occur with fewer observations, which might increase the risk of overfitting. The `scikit-learn` package, by default sets it to 2, thus a range from 2 to 10 was searched over to find the optimal value for this variable.

4. **Minimum Samples per Leaf:** Sets the minimum number of samples that must be present in a leaf node, ensuring that each final node has enough data

to make reliable predictions. While recommended values for this parameter are difficult to find in the literature, package libraries such as `scikit-learn` default to 2. Hence, values from 1 to 10 were searched over.

5. **Maximum Features:** Controls the number of features considered for splitting at each node. By using a subset of features, this parameter injects randomness into the model, creating diverse trees that reduce correlation between them and improve overall performance. This randomness also ensures that the model does not become too reliant on any particular set of features. The literature does not suggest a single optimal *Maximum Features* value exists that fits all datasets, but industry defaults (Han & Kim, 2021) include allowing it to be:
   - $\sqrt{n}$, where n is the number of variables
   - $\log_2(n) + 1$
   - n (using all features)

   These are the approaches we have implemented in our analysis.

By tuning these parameters appropriately, one can effectively balance the trade-off between bias and variance, leading to a more robust and accurate Random Forest model.

Each Random Forest model created from the permutations of values of the previously mentioned hyperparameters is initialized and 5 fold cross-validation is applied. This helps in identifying the best combination of hyperparameters. The grid search is fitted on the training data, and the best parameters are extracted based on the ones which achieved the smallest validation error. The best model on the grid is then retrained on the entire training set. Predictions are made on the test set, and the model's performance is evaluated using mean absolute error (MAE) and mean squared error (MSE). These metrics provide a quantitative measure of the model's accuracy.

5.4.4 Extreme-Gradient Boosted Trees

The process of selecting the best model using XGBoost involved several key steps to ensure the optimal model configuration was discovered. After data preparation, a hyperparameter search space was defined using the `hyperopt` (Bergstra et al., 2013) library in Python. The objective function was set to minimize the negative mean squared error (MSE) using 5-fold cross-validation on the training set. Negative MSE was used to align with the optimization logic of maximizing the objective function.

The Tree-structured Parzen Estimator (TPE) (Watanabe, 2023) algorithm was used to find the best hyperparameters, utilizing 5-fold cross-validation. The following

hyperparameters were optimized during the model selection process over optimal ranges found by Putatunda and Rama (2020):

1. **Boosting Rounds:** Determines the number of trees created during training. There is a bias-variance tradeoff, as increasing the number of rounds reduces bias but increases variance, thus risking overfitting. Conversely, fewer rounds reduce overfitting but increase bias. A total of 200 boosting rounds was found to achieve the right balance, with early stopping rounds of 10 ensuring that training stopped if validation error did not decrease for 10 consecutive rounds. Additionally, once the best model was identified through cross-validation, it was retrained on the full dataset using 20 boosted rounds fewer to further mitigate overfitting. The reduction of 20 boosting rounds was determined through empirical analysis of the validation loss curve. This specific value was selected as it provided a meaningful decrease in model complexity while maintaining performance metrics within acceptable thresholds, thus offering additional protection against overfitting.

2. **Eta (Learning Rate):** Controls the step size during model updates. A smaller value makes the model more robust to overfitting but requires more boosting rounds. The search range for this parameter was set between $\log(1e^{-7})$ and $\log(1)$.

3. **Max Depth:** Determines the maximum depth of each tree. Deeper trees allow the model to capture more complex patterns but increase the risk of overfitting. The search space ranged from 1 to 10.

4. **Subsample:** Represents the fraction of training data randomly sampled for each tree. Lower values introduce additional randomness, helping prevent overfitting. The search range was set between 0.2 and 1.

5. **Colsample bytree:** Specifies the fraction of features randomly sampled for each tree. Like `subsample`, this introduces randomness into the model to reduce overfitting. The search range was set between 0.2 and 1.

6. **Colsample bylevel:** Similar to `colsample bytree`, this parameter controls the fraction of features sampled at each tree level. The search range was set between 0.2 and 1.

7. **Min Child Weight:** Specifies the minimum sum of instance weights (Hessian values) needed in a child node. This parameter plays a crucial role in preventing the model from learning relations that may be highly specific to the training set, thereby reducing overfitting.

    - A **higher** value forces child nodes to have a greater sum of instance

weights, preventing splits that might be due to noise or minor variations in the data.

- A **lower** value allows smaller partitions, potentially capturing intricate patterns but also increasing the risk of overfitting.

A well-chosen value ensures that the model does not create overly specific splits based on a small number of observations, improving generalization. The search space for this parameter was set between $\log(1e^{-16})$ and $\log(e^5)$.

8. **Alpha (L1 Regularization Term):** Adds a penalty equal to the absolute magnitude of coefficients, helping to reduce overfitting by shrinking certain coefficients to zero, effectively performing variable selection. The search space included 0 and values between $\log(1e^{-16})$ and $\log(e^2)$.
9. **Lambda (L2 Regularization Term):** Adds a penalty equal to the square of the magnitude of coefficients, reducing overfitting by preventing extreme coefficient values. The search space included 0 and values between $\log(1e^{-16})$ and $\log(e^2)$.

Figure 30 illustrates the cross-validation and training performance of the optimal model across different numbers of boosting rounds. The model's training was halted early to prevent overfitting, a technique known as *early stopping*. This ensures that training continues only as long as the cross-validation mean absolute error (MAE) improves over each 10-iteration period; otherwise, training is stopped, and the optimal model parameters are selected. If the number of boosting rounds were increased indefinitely, we would expect the training error to continue decreasing while the cross-validation error would likely increase due to overfitting. To determine the optimal number of boosting rounds for final model training, we selected the number of rounds where the cross-validation MAE stopped improving and used a value 20 rounds earlier. This approach further mitigated the risk of overfitting. The optimal model configuration was then initialized, and the model was re-trained with the specified number of earlier boosting rounds on the full dataset. The model's performance was then evaluated on the test set using MSE and MAE.

5.4.5 Autoencoder

To optimize the autoencoder used on each dataset of explanatory variables, numerous rounds of testing were implemented. Functions were defined to train multiple autoencoder models for a specified middle layer size (bottleneck). Thus, for data with 20 variables, a series of autoencoders were built with bottlenecks of 5, 10, and 15 neurons, respectively. For each bottleneck, the optimal L2-regularization values were found, as well as the optimal node ratio value. However, in this case, the output

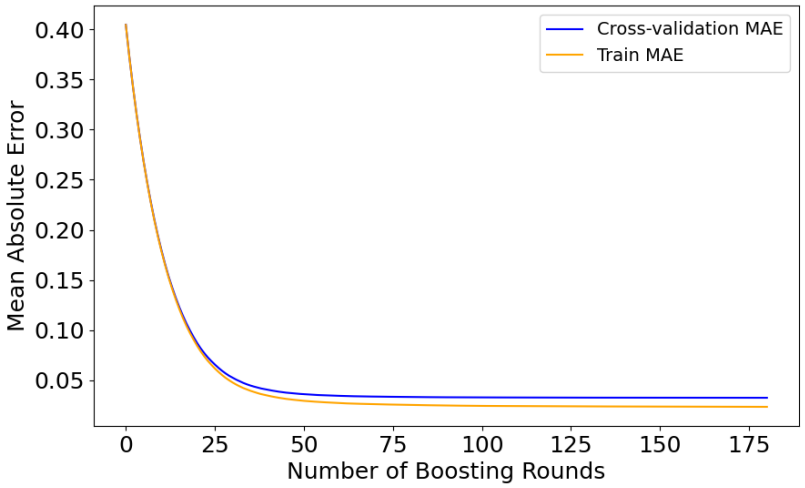

Figure 30: Optimal XGBoost model cross-validation and training MAE over Boosting Rounds

dimension of the autoencoder was the bottleneck size, and the encoder and decoder had the same size. The descriptions of all the hyperparameters are shown below:

1. **Ratio:** The ratio parameter determines the rate at which data is compressed through the autoencoder's layers before reaching its minimal representation. A higher ratio creates fewer, more aggressive compression layers, while a lower ratio leads to more gradual compression across more layers. Once again, the optimal ratio value was selected from a sequence of values between 0.1 and 0.9.

2. **Activation Functions:** The activation functions experimented with included, the Identity function, Sigmoid function, LeakyReLU function, ReLU function, and Tanh function.

Once the models were created, they were grouped based on common bottleneck sizes (number of columns after compression ie. middle layer size) to find the optimal model for each bottleneck size. After splitting the data into 70-15-15 training, validation, and test sets, each model was trained, and its validation performance recorded. For each of the bottleneck sizes (75%, 50%, and 25% of the original feature dimensionality), we identified the optimal autoencoder architecture through systematic evaluation. Specifically, for each bottleneck size, numerous hyperparameter combinations were tested, including different activation functions, ratios for determining hidden layer sizes, and a range of regularization values (λ) were applied

during training. For each combination of these hyperparameters, an autoencoder model was generated and trained. The models were trained using Mean Squared Error (MSE) as the loss function, with the Adam optimizer employed to minimize this loss while incorporating L2-regularization.

The average validation loss over the last 100 epochs was continually calculated to assess the stability and effectiveness of each model configuration. The model configuration with the lowest validation loss was identified as the best model. This optimal model configuration was then finally re-trained across multiple λ values to prevent overfitting.

The goal of the above process was to select the most suitable autoencoder architecture for the task at hand, balancing model complexity and generalizability. The selected model aimed to compress the data into a reduced format that best represented the underlying relationships in the data without overfitting to the training data.

5.5 App Development

This section outlines the development of the football DSS designed using the Python package Streamlit (Developers, 2024) to showcase the findings of the *OBV* models in a way that would provide decision support to prospective scouters and others involved in player recruitment. Specifically, the application comprises an easy-to-use interface that allows users to upload raw data from Wyscout, and output the predicted *Pass OBV*, *DC OBV*, *DA OBV*, and *Total OBV* per player. The system's design was informed by Jakob Nielsen's usability heuristics described by Mirkowicz and Grodner (2018), to promote a seamless and easy user experience.

5.5.1 Data Collection and Preprocessing

Data Sources: Player performance data can be downloaded from Wyscout and uploaded straight to the app in seconds. This data contains all the necessary variables for the models to predict the corresponding *OBV* components.

Data Cleaning: Once the data is uploaded into the app, and the "Generate OBV" button is clicked, the data undergoes the following cleaning processes:

- Goalkeepers are removed from the dataset.
- Players with less than 500 minutes of playtime are filtered out to ensure the ratings are robust to time. Players with less playtime are less likely to maintain consistent *OBV* scores over a longer period.

- Player positions are categorized into five main categories: Fullbacks, Centerbacks, Midfielders, Forwards, and Wingers.

The dataset is then scaled, with one-hot encoding applied to the position categories. The full Wyscout dataset is subset for each model.

5.5.2 Visualization

Upon inserting the Wyscout data, predictions of *Total OBV*, *Pass OBV*, *DC OBV*, and *DA OBV* are generated. The predictions are output to a '.xlsx' file with conditional formatting applied to highlight key performance indicators.

The DSS provides a customizable visualization interface that enables scouts to generate comparative scatterplots of OBV components, facilitating intuitive analysis of players' in-game attributes. Market valuations sourced from Wyscout are also integrated into the analysis framework, allowing for direct comparison between a player's performance metrics and their market value. This integration is further enhanced by a feature-engineered metric, *OBV* per million, which quantifies performance relative to market value, enabling scouts to identify players who offer optimal value in the transfer market.

5.6 Conclusion

The methods detailed in this section establishes a robust approach to investigate the feasibility of modeling *OBV* using readily available frequency statistics. The process begins with a foundational phase of EDA employing techniques such as PCA and t-SNE to uncover underlying data structures. Dimensionality reduction—through PCA, Autoencoders, and variable importance metrics then ensue to create optimized feature sets for modeling.

A suite of predictive models, spanning from classical linear regression (OLS) to advanced ensemble (Random Forest, XGBoost) and neural network-based methods (MLP, Autoencoder), will be comparatively evaluated. Each model is subjected to rigorous optimization processes, involving iterative hyperparameter tuning, cross-validation, and systematic variable selection strategies to ensure peak performance and prevent overfitting. The culmination of these methods is not only a set of predictive models but also the development of a practical DSS, designed to translate complex model outputs into actionable insights for a scouting department. Collectively, these procedures provide a comprehensive framework for evaluating the central hypothesis and delivering a tangible, low-cost solution for player evaluation.

6 Results

For conciseness, this analysis presents a subset of the modeling results. While numerous models were fitted across the OBV components, the most pertinent findings have been focused on below. Specifically, for each dataset, we present one SHAP (Wang et al., 2024) plot highlighting key feature importances from a selected machine learning model, alongside detailed results from our linear models. The linear model results are emphasized due to their practical advantage in football club settings, where they can be more readily implemented and interpreted compared to complex machine learning models. This selective presentation allows for effective communication of the key insights while demonstrating the models' capabilities across different datasets.

The models presented in this section highlight the capabilities of different approaches to modeling OBV and its components. From using different model building strategies to implementing different variable selection methods, the work shown below offers a wide-range of information regarding optimal modeling strategies dependent on the dataset. For conciseness, model names have been abbreviated; however, Table 6 below provides a detailed list of all relevant models implemented in this dissertation.

Model Name	**Description**
LM Original Model	Linear model with all variables incorporated
LM Reduced Model	Linear model with top-down AIC variable selection approach
XGB Cumulative Importance Method	eXtreme Gradient Boosting with Cumulative Importance approach
XGB Threshold Method	eXtreme Gradient Boosting with top N most important features selected based on a predefined importance threshold
XGB Top N Features Method	eXtreme Gradient Boosting with top N most important features selected
XGB Autoencoder (1 quarter)	eXtreme Gradient Boosting with autoencoder-compressed input to one-quarter of the original input size
XGB Autoencoder (2 quarter)	eXtreme Gradient Boosting with autoencoder-compressed input to half of the original input size

Continued on next page

Model Name	**Description**
XGB Autoencoder (3 quarter)	eXtreme Gradient Boosting with autoencoder-compressed input to three-quarters of the original input size
XGB PCA 80% Method	eXtreme Gradient Boosting with PCA input that accumulatively accounts for just over 80% of the variation
XGB PCA 90% Method	eXtreme Gradient Boosting with PCA input that accumulatively accounts for just over 90% of the variation
RF Cumulative Importance Method	Random Forest with Cumulative Importance approach
RF Threshold Method	Random Forest with top N most important features selected based on a predefined importance threshold
RF Top N Features Method	Random Forest with top N most important features selected
RF Autoencoder (1 quarter)	Random Forest with autoencoder-compressed input to one-quarter of the original input size
RF Autoencoder (2 quarter)	Random Forest with autoencoder-compressed input to half of the original input size
RF Autoencoder (3 quarter)	Random Forest with autoencoder-compressed input to three-quarters of the original input size
RF PCA 80% Method	Random Forest with PCA input that accumulatively accounts for just over 80% of the variation
RF PCA 90% Method	Random Forest with PCA input that accumulatively accounts for just over 90% of the variation
MLP Original Model	MLP model with all available input features

Table 6: List of models with their descriptions

6.1 Total OBV

6.1.1 Wyscout Dataset

6.1.1.1 Top five models

Table 7 below shows the results of the top 5 performing models predicting *Total OBV* when trained on the Wyscout dataset.

Metric	LM Original	LM Reduced	MLP Original	XGB Cumulative	XGB Threshold
Test Set MAE	0.0476	0.0476	0.0487	0.0491	0.0496
Test Set MSE	0.0036	0.0036	0.0040	0.0039	0.0039
Precision@5 CB	0.4	0.2	0.3	0.4	0.4
Precision@5 FB	0.2	0.2	0.2	0.6	0.4
Precision@5 M	0.6	0.6	0.4	0.4	0.4
Precision@5 W	0.8	0.8	0.4	0.2	0.2
Precision@5 F	0.6	0.6	0.6	0.4	0.4
Avg Precision@5	0.52	0.48	0.38	0.4	0.36
Precision@10 CB	0.5	0.5	0.6	0.6	0.5
Precision@10 FB	0.6	0.6	0.7	0.6	0.6
Precision@10 M	0.4	0.4	0.5	0.5	0.5
Precision@10 W	0.7	0.7	0.6	0.5	0.4
Precision@10 F	0.7	0.7	0.7	0.6	0.6
Avg Precision@10	0.58	0.58	0.62	0.56	0.52

Table 7: Summary of top-performing models' performance metrics when modeling *Total OBV* on the Wyscout dataset

Table 7 above reveals the Linear Models (LM Original and LM Reduced) demonstrate superior performance with the lowest MAE and MSE scores when compared to the machine learning models. The reduced linear model achieves a lower AIC value (-2363.2) when compared to the original linear model (-2308.63), showing the reduction in variables helps balance the predictive accuracy and parsimonious nature of the model better. Given that the difference between the first quantile and the third quantile of the *Total OBV* distribution was 0.13, an MAE of 0.0476 showed moderate modeling capabilities. Precision@K varied significantly across player positions, with forwards and wingers consistently being predicted accurately, while model's generally struggled to accurately order centre backs often precision. For wingers specifically, the linear models demonstrated strong predictive power by successfully capturing 7 of the top 10 players in their predicted top 10 rankings. The

XGB models showed competitive performance when compared to the linear models, particularly in terms of Precision@K for centre backs and full backs.

6.1.1.2 Influence of features on model output

A useful tool for visualising the contribution of each feature to a models prediction is a SHAP (SHapeley Additive exPlanations) plot (Lundberg & Lee, 2017). The SHAP plot in Figure 31 below indicates the effects variables have on the predicted *Total OBV* for the original multilayer perceptron model (MLP Original Model).

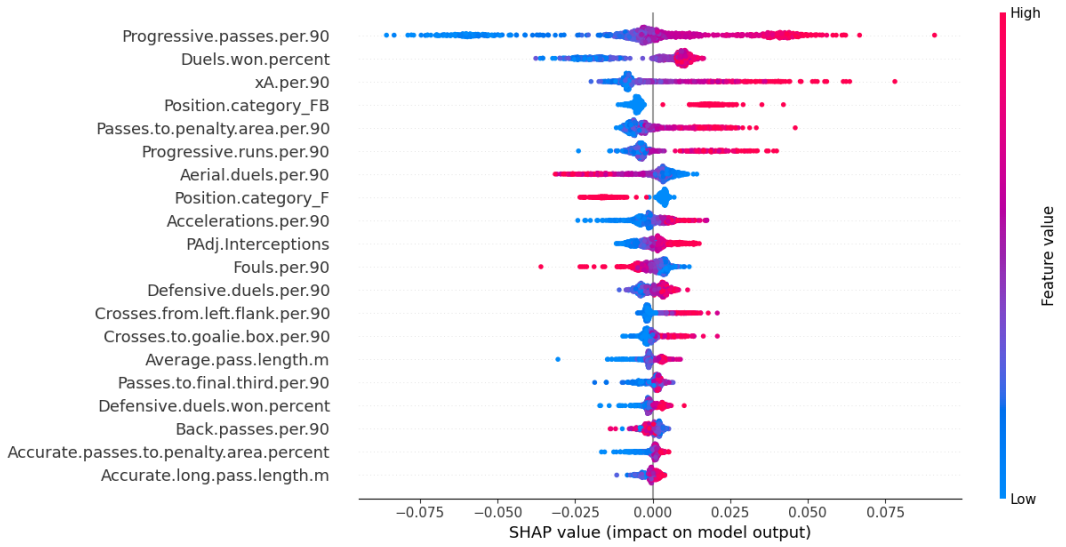

Figure 31: SHAP values for the MLP model fitted to the full Wyscout dataset

Progressive.passes.per.90, *xA.per.90* (expected assists), as well as whether the player is a fullback or not all had a significant effect on the predictions made by the MLP, with *Progressive.passes.per.90* having the largest effect. Progressive actions, as expected, increased the predicted *OBV*, while the effect of the percentage completed metrics was less significant. Being a fullback had a positive effect on the predicted *OBV*, while Forwards were hampered by their positioning.

6.1.1.3 Top-performing linear model

Table 8 below details the reduced linear model used in predicting *Total OBV* on the Wyscout dataset.*Progressive.passes.per.90* emerges as one of the most influential features for predicting *OBV*, with the highest positive coefficient, and a small p-value (≤ 0.001). Of note is that this model contains many variables that do not reach statistical significance. The model explains a substantial portion of the variance,

with an Adjusted R-squared value of 0.709. Numerous position-specific interactions show significant effects, suggesting that the impact of certain actions varies by player position.

Variable	**Beta**	**95% CI**	**p-value**
(Intercept)	0.120	0.108, 0.132	≤ 0.001
FB	0.036	0.016, 0.056	≤ 0.001
xA per 90	0.033	0.027, 0.039	≤ 0.001
Progressive passes per 90	0.032	0.021, 0.044	≤ 0.001
Duels won percent	0.032	0.021, 0.043	≤ 0.001
M	0.026	0.008, 0.044	0.004
W	0.022	0.002, 0.041	0.033
W*Progressive runs per 90	0.020	0.003, 0.037	0.023
Accelerations per 90	0.018	0.003, 0.033	0.017
FB*Progressive passes per 90	0.017	0.003, 0.031	0.015
M*PAdj Interceptions	0.017	0.005, 0.028	0.005
Progressive runs per 90	0.012	0.001, 0.022	0.031
Passes to penalty area per 90	0.010	0.002, 0.019	0.022
Defensive duels per 90	0.009	0.003, 0.015	0.003
Average pass length m	0.007	0.000, 0.014	0.037
PAdj Sliding tackles	0.005	0.001, 0.010	0.028
Crosses to goalie box per 90	0.005	-0.002, 0.011	0.156
Accurate passes to penalty area percent	0.004	-0.000, 0.008	0.062
Accurate back passes percent	0.004	-0.000, 0.008	0.069
Accurate crosses from left flank percent	0.003	-0.001, 0.008	0.094
Successful dribbles percent	-0.004	-0.008, 0.000	0.070
Defensive duels won percent	-0.006	-0.012, 0.000	0.056
Smart passes per 90	-0.006	-0.012, -0.000	0.040
Passes to final third per 90	-0.007	-0.017, 0.003	0.148
Back passes per 90	-0.008	-0.015, -0.002	0.013
Aerial duels won percent	-0.009	-0.015, -0.002	0.008
Aerial duels per 90	-0.010	-0.016, -0.005	≤ 0.001
Fouls per 90	-0.011	-0.017, -0.006	≤ 0.001
M*Accelerations per 90	-0.015	-0.030, 0.001	0.074
FB*Accelerations per 90	-0.017	-0.033, -0.001	0.043
Progressive runs per 90*CB	-0.026	-0.044, -0.008	0.005
W*Accelerations per 90	-0.035	-0.054, -0.015	≤ 0.001

Table 8: Model summary of the optimal linear model, sorted by decreasing order of beta coefficients, with 95% Confidence Intervals (CI) and p-values

6.1.1.4 Top performing linear model: residual diagnostics

Figure 32 below shows the residual diagnostic outputs from the reduced linear model.

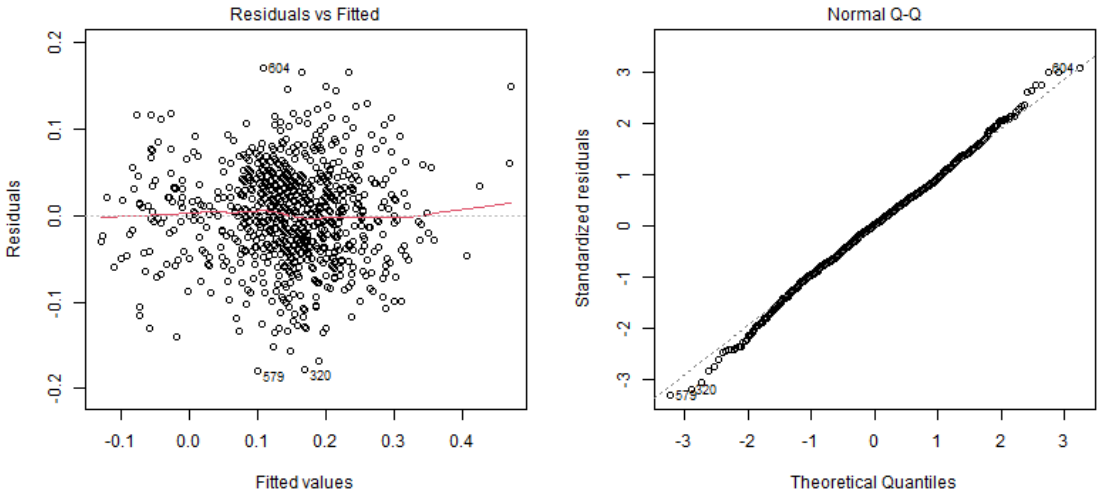

Figure 32: Linear Model Residuals vs Fitted and Q-Q plot of the model's residuals

The Residuals vs Fitted plot shows a relatively even spread around zero. Heteroscedasticity does appear to be present however, as the Residuals vs Fitted plot is somewhat funnel shaped, with variance increasing for higher fitted values. A Breusch-Pagan test for heteroscedacity (Breusch & Pagan, 1979) was conducted and a p-value of 0.004 was obtained, thus confirming that there is strong evidence for the presence of heteroscedacity in the data. The Q-Q plot suggests a near-normal distribution of residuals with marginal deviation at the extremes and is supported by a p-value of 0.16 from the Shapiro-Wilks test for normality (Shapiro & Wilk, 1965).

6.1.2 FBref Dataset

6.1.2.1 Top five models

Table 9 below shows the *Total OBV* modeling results for the FBref dataset. The linear models again demonstrated the best performance in terms of MAE and MSE, consistent with the Wyscout dataset results. Interestingly, although the original linear model achieves better MAE and MSE scores, it achieves a larger AIC value (-879.04) than the simpler linear model (-917.64). Precision@K scores are relatively

high when compared to the Wyscout dataset, with numerous perfect scores for wingers and forwards across all top five performing models, suggesting these models were able to pick the correct top 10 wingers and forwards. The average Precision@5 values are the same for the linear models, with the average Precision@10 value for the reduced linear model being slightly higher than the original linear model.

Metric	LM Original	LM Reduced	RF Threshold	RF Cumulative	XGB Original
Test Set MAE	0.0492	0.0499	0.0500	0.0500	0.0517
Test Set MSE	0.0037	0.0038	0.0048	0.0048	0.0049
Precision@5 CB	0.4	0.6	0.6	0.6	0.2
Precision@5 FB	0.8	0.8	0.4	0.4	0.4
Precision@5 M	0.8	0.8	0.6	0.6	0.4
Precision@5 W	1.0	1.0	0.8	0.8	0.6
Precision@5 F	0.8	0.8	0.6	0.6	0.6
Avg Precision@5	0.76	0.80	0.60	0.60	0.44
Precision@10 CB	0.9	0.9	0.6	0.6	0.6
Precision@10 FB	0.6	0.6	0.9	0.9	0.9
Precision@10 M	0.7	0.8	0.8	0.8	0.7
Precision@10 W	1.0	1.0	0.8	0.8	0.7
Precision@10 F	1.0	1.0	1.0	1.0	1.0
Avg Precision@10	0.84	0.86	0.82	0.82	0.78

Table 9: Summary of top-performing models' performance metrics modeling *Total OBV* on the FBref dataset

6.1.2.2 Influence of features on model output

Figure 33 below shows the optimal Random Forest model's SHAP values. Clearly, progressive actions and passes near the opposition's goal have a large effect on the predicted *OBV* value, such as *Progressive.passing.distance.per.90*, *Progressive.carries.distance.per.90*, *Passes.into.penalty.area*, and *Crosses.into.penalty.area*. Defensive actions such as *Interceptions.per.90* and *Tackles.in.defensive.3rd.per.90* can increase the predicted *OBV* if high enough. The most significant dribbling attribute in predicting *Total OBV* is *Progressive.carries.distance.per.90*.

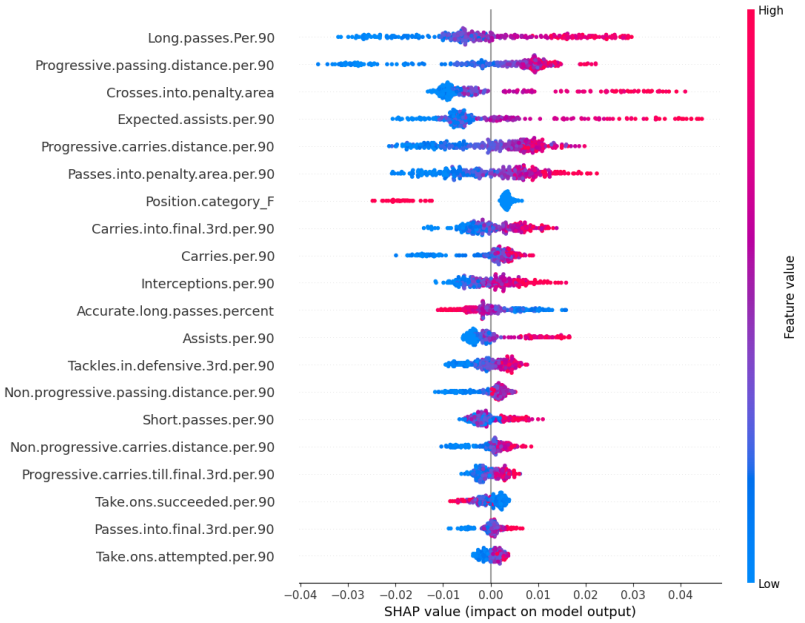

Figure 33: SHAP values for Random Forest model fitted to full FBref dataset

6.1.2.3 Top-performing linear model

Table 10 below summarises the reduced linear model used when modeling OBV using the FBref dataset. Position-specific interactions with progressive actions (passes and carries) showed statistical significance, highlighting the differing value of these skills across the different positions, and the general importance of actions which progress the ball. The model demonstrated a higher Adjusted R-squared (0.734) compared to the Wyscout dataset model, indicating better overall fit. The presence of both positive and negative coefficients for similar actions across different positions showed the importance of where that action occurs on the pitch.

Variable	Beta	95% CI	p-value
(Intercept)	0.167	0.157, 0.176	≤ 0.001
Progressive passing distance per 90*M	0.065	0.021, 0.110	0.004
Progressive carries distance per 90	0.059	0.026, 0.092	≤ 0.001
W*Non progressive carries distance per 90	0.055	0.018, 0.092	0.004
Non progressive carries distance per 90*CB	0.049	0.012, 0.086	0.011
Long passes Per 90	0.038	0.020, 0.056	≤ 0.001
Progressive passing distance per 90	0.036	0.013, 0.060	0.003
Non progressive carries distance per 90*M	0.035	0.002, 0.069	0.038
Passes into penalty area per 90	0.028	0.016, 0.040	≤ 0.001
Interceptions per 90	0.027	0.015, 0.039	≤ 0.001
Expected assists per 90	0.020	0.004, 0.036	0.017
Assists per 90	0.018	0.001, 0.035	0.040
Carries into final 3rd per 90	0.018	0.005, 0.031	0.008
Accurate short passes percent	0.016	0.006, 0.026	0.002
Carries into penalty area per 90	0.014	0.002, 0.026	0.019
Tackles in defensive 3rd per 90	0.010	0.002, 0.019	0.014
Tackles in midfield 3rd per 90	0.007	-0.001, 0.014	0.078
Take ons succeeded per 90	−0.004	-0.010, 0.002	0.158
Assists Minus Expected Assisted goals per 90	−0.010	-0.022, 0.002	0.096
Accurate long passes percent	−0.011	-0.019, -0.004	0.004
Through balls per 90	−0.013	-0.022, -0.004	0.006
Interceptions per 90*M	−0.013	-0.031, 0.005	0.142
Short passes per 90	−0.019	-0.037, -0.001	0.035
Carries per 90	−0.026	-0.057, 0.004	0.094
Interceptions per 90*FB	−0.028	-0.051, -0.005	0.016
FB*Take ons attempted per 90	−0.029	-0.060, 0.001	0.056
Non progressive carries distance per 90	−0.042	-0.073, -0.010	0.010
Progressive carries distance per 90*CB	−0.043	-0.078, -0.008	0.016
W*Progressive carries distance per 90	−0.043	-0.083, -0.004	0.032
Long passes Per 90*CB	−0.043	-0.066, -0.020	≤ 0.001
W*Progressive passing distance per 90	−0.046	-0.107, 0.014	0.132
Long passes Per 90*M	−0.050	-0.079, -0.021	≤ 0.001
Progressive carries distance per 90*M	−0.053	-0.093, -0.014	0.008
W	−0.058	-0.114, -0.003	0.041

Table 10: Regression results sorted by Beta coefficients, with 95% Confidence Interval (CI) and p-values.

6.1.2.4 Top performing linear model: residual diagnostics

Figure 34 below shows the residual diagnostics for the reduced linear model on the FBref dataset.

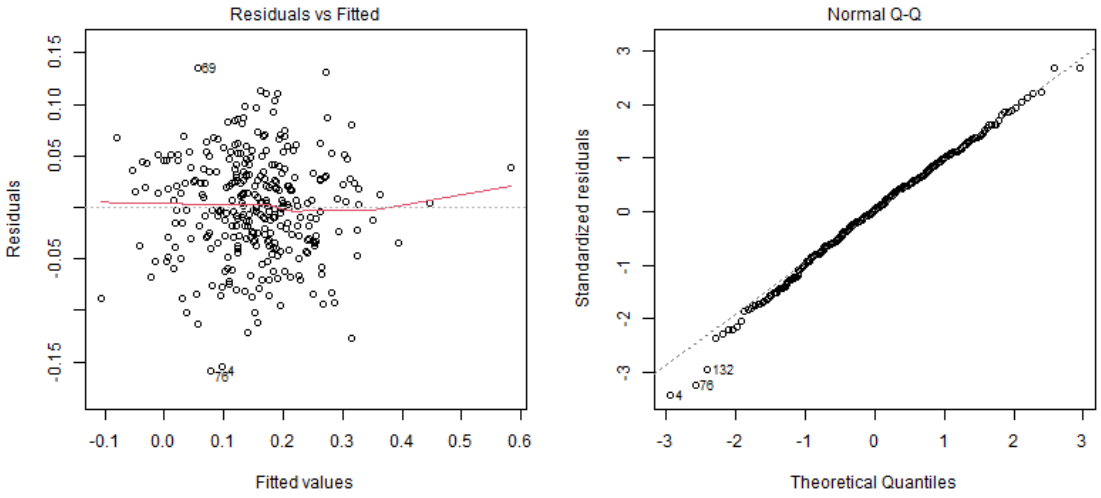

Figure 34: Linear Model fitted to *Total OBV* FBref dataset Diagnostic Plots

The Residuals vs Fitted plot shows a similar scatter to the best-performing linear model on the Wyscout dataset, suggesting non-linear methods are not necessary in exploring these relationships. A smaller level of heteroskedasticity seems prevalent, suggesting a more consistent accuracy across the range of fitted values. This is further confirmed by a p-value of 0.3 in a Breusch-Pagan test. The Q-Q plot suggests the distribution of the residuals is approximately normal confirmed by a Shapiro-Wilks test p-value of 0.59.

6.2 Pass OBV

6.2.1 Wyscout Dataset

6.2.1.1 Top five models

Table 11 below shows the performance of the top 5 models trained on the Wyscout passing dataset to predict *Pass OBV*.

Metric	LM Original	LM Reduced	XGB Cumulative	MLP Original	XGB Original
Test Set MAE	0.0273	0.0276	0.0320	0.0322	0.0324
Test Set MSE	0.0013	0.0013	0.0018	0.0016	0.0018
Precision@5 CB	0.4	0.4	0.6	0.2	0.2
Precision@5 FB	0.2	0.4	0.4	0.4	0.4
Precision@5 M	0.6	0.6	0.4	0.2	0.2
Precision@5 W	0.6	0.6	0.4	0.2	0.2
Precision@5 F	0.6	0.6	0.4	0.4	0.4
Avg Precision@5	0.48	0.52	0.44	0.28	0.28
Precision@10 CB	0.7	0.6	0.6	0.7	0.7
Precision@10 FB	0.6	0.6	0.6	0.6	0.6
Precision@10 M	0.6	0.6	0.6	0.7	0.7
Precision@10 W	0.8	0.8	0.8	0.7	0.7
Precision@10 F	0.6	0.6	0.5	0.6	0.6
Avg Precision@10	0.66	0.64	0.62	0.66	0.66

Table 11: Summary of top-performing models' performance metrics when modeling *Pass OBV* on the Wyscout passing dataset

The optimal test set MAE of 0.0273 for *Pass OBV* indicates strong predictive performance, considering that the inter-quartile range for *Pass OBV* is 0.082. Linear Models again outperform other more complex ML models, with approximately 12% lower MAE when moving from the XGB Cumulative model to the linear models. The top linear models once again achieve similar performance, however the reduced linear model also has a significantly lower AIC value of -3017.92 when compared to the original models value of -2990.14. Precision scores are generally good for the Precision@10 score across all models, with values ranging from 0.5 to 1, while Precision@5 has values as low as 0.2, indicating that only 1 out of the top 5 players was correctly predicted in the top 5 for a particular position.

6.2.1.2 Influence of features on model output

SHAP values for the XGB Cumulative model are shown in Figure 35 below.

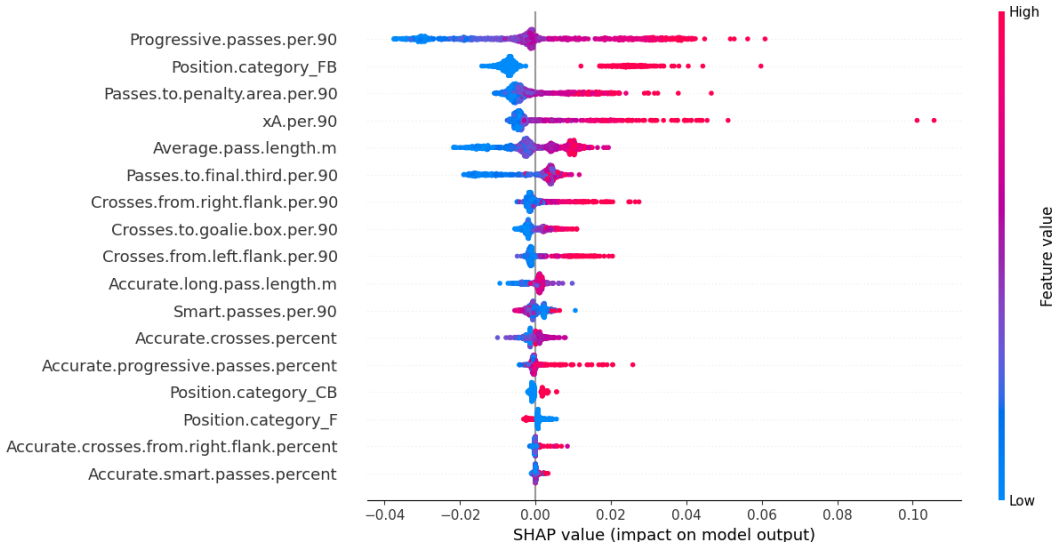

Figure 35: SHAP values for XGB model fitted to the Pass Wyscout dataset

The most important variable from the passing Wyscout dataset was still *Progressive.passes.per.90*. *Average.pass.length.m* was also important, with the longer the average pass the better the predicted *Pass OBV*. The number of crosses also held some predictive power, while the accuracy metrics once again held some of the least importance in predicting *Pass OBV*. Player position was once again important, with fullbacks being the most important.

6.2.1.3 Top-performing linear model

Table 12 below describes the reduced linear model. The Adjusted R^2 value was improved when predicting only *Pass OBV* (0.758) compared to predicting *Total OBV* (0.709). *Progressive.passes.per.90* remained the most influential feature for *Pass OBV*, reinforcing its importance in player valuation. Position-specific interactions with progressive passes showed significant negative coefficients, suggesting that the default position (forwards) have the highest *Pass OBV* reward for a progressive pass. Crosses were also found to be highly significant, and accuracy metrics were found to be least informative of *Pass OBV*.

Variable	Beta	95% CI	p-value
(Intercept)	0.054	0.049, 0.059	≤ 0.001
Progressive passes per 90	0.059	0.052, 0.065	≤ 0.001
FB	0.034	0.023, 0.045	≤ 0.001
xA per 90	0.033	0.029, 0.037	≤ 0.001
CB	0.024	0.012, 0.035	≤ 0.001
FB * Accurate long pass length (m)	0.015	0.002, 0.027	0.021
Crosses from right flank per 90	0.009	0.005, 0.012	≤ 0.001
Crosses from left flank per 90	0.008	0.004, 0.011	≤ 0.001
Accurate crosses from left flank percent	0.003	0, 0.005	0.065
Accurate back passes percent	0.003	0, 0.006	0.060
Accurate progressive passes percent	0.002	-0.001, 0.006	0.14
Smart passes per 90	-0.006	-0.01, -0.002	0.002
Short medium passes per 90	-0.006	-0.011, -0.001	0.015
Back passes per 90	-0.009	-0.014, -0.003	0.001
FB * Progressive passes per 90	-0.011	-0.021, 0	0.046
Progressive passes per 90 * M	-0.023	-0.032, -0.015	≤ 0.001
CB * Progressive passes per 90	-0.030	-0.041, -0.019	≤ 0.001

Table 12: Regression results sorted by Beta coefficients, with 95% Confidence Interval (CI), p-values.

6.2.1.4 Top performing linear model: residual diagnostics

Figure 36 below shows the residual diagnostic plots for the reduced linear model on the Wyscout passing dataset. While the Residuals vs Fitted plot appeared to have shown a scatter of residuals over the fitted values, a Breusch-Pagan test for heteroskedasticity obtained a p-value of 2.62×10^{-11}, leading to a rejection of the null hypothesis of homoskedasticity. The Q-Q plot indicated that the distribution exhibited good normality at its center but showed significant deviation in the tails. This observation is consistent with the Shapiro-Wilk test, which returned a p-value of 6.545×10^{11}.

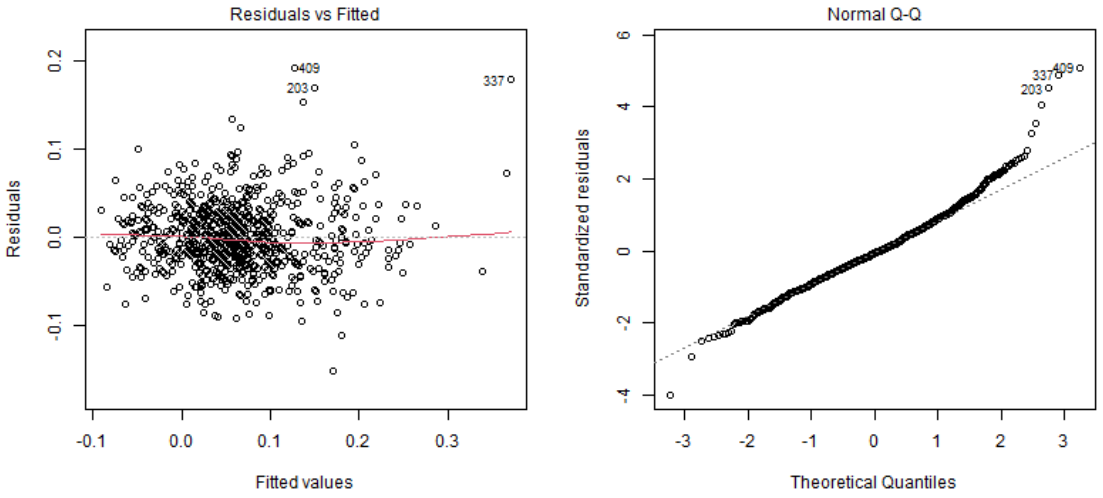

Figure 36: Linear Model on *Pass OBV* Wyscout dataset Diagnostic Plots

6.2.2 FBref Dataset

6.2.2.1 Top five models

Table 13 below shows the top 5 model performances for models trained on the FBref passing dataset. Further reduced MAE test set values were observed, with the optimal model achieving a value of 0.0239. Linear Models achieve the lowest MAE and MSE again, consistent with other *OBV* predictions and reinforcing their effectiveness for this specific task. Although the original linear model slightly outperforms the reduced model based on MAE and MSE, their AIC scores once again favour the reduced model (-917.64) over the original model (-879.04). Precision@K scores are notably higher for the FBref passing dataset compared to the corresponding Wyscout dataset, particularly for fullbacks and midfielders, suggesting that the FBref features may better capture passing performance for these positions. High Precision@K scores across multiple model types were found, with perfect scores being achieved for forwards for Precision@10, implying all 10 of the highest performing forwards (with respect to *Pass OBV*) were predicted to be in the top 10, and a lowest score across all models and positions of 0.6.

Metric	LM Original	LM Reduced	RF Top N	XGB Top N	RF Original
Test Set MAE	0.0239	0.0246	0.0325	0.0334	0.0334
Test Set MSE	0.0014	0.0014	0.0028	0.0021	0.0028
Precision@5 CB	0.4	0.6	0.4	0.2	0.4
Precision@5 FB	0.8	0.8	0.8	0.6	1.0
Precision@5 M	0.8	0.8	0.6	0.8	0.8
Precision@5 W	0.4	0.4	0.8	0.6	0.8
Precision@5 F	0.6	0.6	0.6	0.6	0.6
Avg Precision@5	0.60	0.64	0.64	0.56	0.72
Precision@10 CB	0.8	0.7	0.6	0.6	0.7
Precision@10 FB	0.9	0.9	0.9	0.9	0.9
Precision@10 M	0.8	0.8	0.8	0.8	0.9
Precision@10 W	0.9	0.9	0.6	0.6	0.6
Precision@10 F	1.0	0.9	1.0	1.0	1.0
Avg Precision@10	0.88	0.84	0.78	0.78	0.82

Table 13: Summary of top-performing models' performance metrics modeling *Pass OBV* on the FBref pass dataset

6.2.2.2 Influence of features on model output

SHAP values for the top-performing RF model are shown in Figure 37 below.

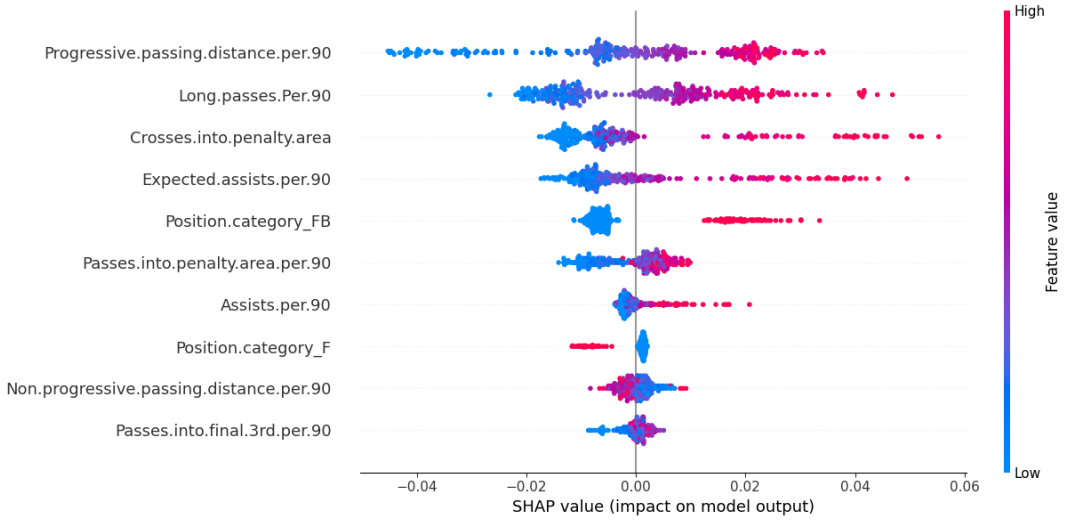

Figure 37: SHAP plot for Random Forest model fitted to Pass FBref dataset

Progressive.Passing.Distance.per.90 and *Expected.Assists.per.90* are once again deemed important when predicting *Pass OBV*, as well as *Long.Passes.per.90* and *Crosses.into.penalty.area*. Neither *Passes.into.final.3rd.per.90* of the pitch nor *Passes.into.penalty.area.per.90* are deemed as important. Forwards were again hampered by their position being predominantly further up the pitch when it comes to predicting *Pass OBV*, as being a forward is seen to slightly decrease a players predicted *Pass OBV*.

6.2.2.3 Top-performing linear model

Table 14 below shows the reduced linear model fitted to the FBref passing dataset.

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	0.069	0.060, 0.077	≤ 0.001
Progressive passing distance per 90	0.091	0.075, 0.107	≤ 0.001
Long passes per 90 * FB	0.042	0.029, 0.055	≤ 0.001
Expected assists per 90	0.037	0.030, 0.044	≤ 0.001
Long passes per 90 * W	0.034	0.015, 0.054	≤ 0.001
M * Long passes per 90	0.031	0.018, 0.044	≤ 0.001
Accurate short passes percent	0.007	0.001, 0.012	0.023
Passes into penalty area per 90	0.006	-0.002, 0.013	0.12
Crosses into penalty area	0.005	0, 0.011	0.069
Assists Minus Expected Assisted goals per 90	-0.003	-0.006, 0.001	0.200
Progressive passes received per 90	-0.009	-0.016, -0.001	0.031
Through balls per 90	-0.012	-0.017, -0.007	≤ 0.001
M	-0.025	-0.038, -0.012	≤ 0.001
Non-progressive passing distance per 90	-0.033	-0.041, -0.025	≤ 0.001
Progressive passing distance per 90 * W	-0.036	-0.059, -0.014	0.002
Progressive passing distance per 90 * FB	-0.040	-0.064, -0.016	0.001
Progressive passing distance per 90 * CB	-0.059	-0.077, -0.041	≤ 0.001
M * Progressive passing distance per 90	-0.062	-0.085, -0.038	≤ 0.001

Table 14: Regression results sorted by Beta coefficients, with 95% Confidence Interval (CI) and p-values.

Progressive.passing.distance.per.90 had the largest positive coefficient, consistent with its importance in other models. The model achieved a high Adjusted R-squared of 0.823, indicating that the selected features capture a large portion of the variance and allow for high predictive power for *Pass OBV*. The presence of significant coefficients for various types of passes (e.g.long passes, through balls) provided insights into the specific passing actions that contribute most to a player's value. Only one accuracy metric (Accurate short passes percent) was selected in the final model.

6.2.2.4 Top performing linear model: residual diagnostics

Figure 38 below shows the residual diagnostic plot for the reduced linear model on the FBref passing dataset.

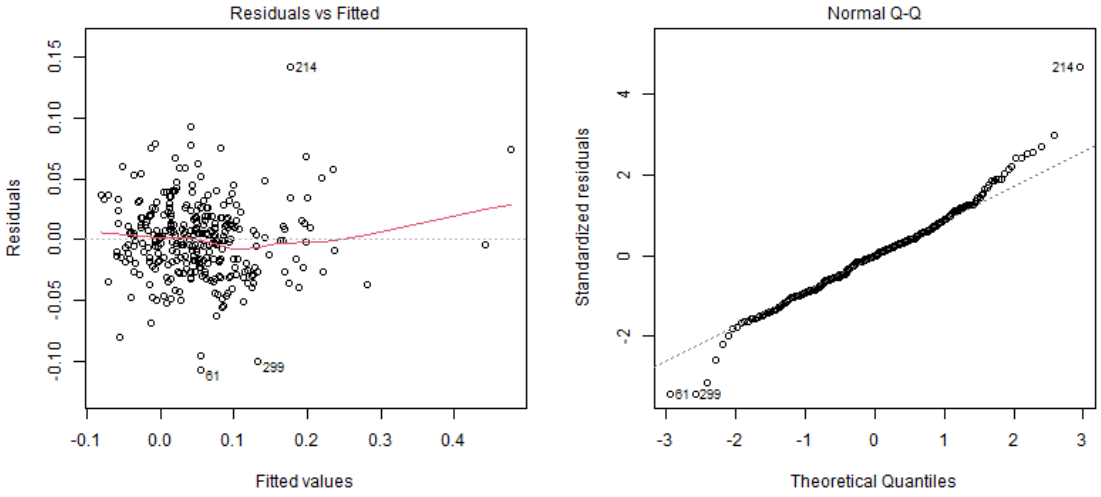

Figure 38: Linear Model on *Pass OBV* FBref dataset Diagnostic Plots

The Residuals vs Fitted plot shows a similar random scatter to that of the best performing linear model on the Wyscout passing dataset, with marginal evidence of heteroskedasticity, with a Breusch-Pagan test p-value of 0.0598 confirming this. This plot suggested that for higher fitted values, the model's predictive accuracy may suffer. The Q-Q plot indicated good normality in the center but deviation from normality in the tails of the distribution, leading to a Shapiro-Wilk test p-value of 0.0002, which strongly indicates that the residuals deviate significantly from normality, leading us to reject the null hypothesis of normality.

6.3 DC OBV

6.3.1 Wyscout Dataset

6.3.1.1 Top five models

Table 15 below shows the top five performing models fitted to the dribbles and carries Wyscout dataset.

Metric	MLP Original	RF Original	RF PCA 99%	RF Threshold	RF Top N	LM Reduced
Test Set MAE	0.0216	0.0219	0.0219	0.0220	0.0222	0.0233
Test Set MSE	0.0009	0.0009	0.0009	0.0009	0.0009	0.0010
Precision@5 CB	0.2	0.2	0.2	0.2	0.2	0.6
Precision@5 FB	0.6	0.6	0.4	0.6	0.6	0.6
Precision@5 M	0.4	0.4	0.6	0.2	0.2	0.4
Precision@5 W	0.4	0.4	0.4	0.6	0.4	0.6
Precision@5 F	0.4	0.4	0.4	0.4	0.4	0.4
Avg Precision@5	0.40	0.40	0.40	0.40	0.36	0.52
Precision@10 CB	0.3	0.3	0.3	0.4	0.4	0.6
Precision@10 FB	0.8	0.8	0.8	0.8	0.8	0.6
Precision@10 M	0.4	0.4	0.5	0.5	0.4	0.5
Precision@10 W	0.6	0.6	0.6	0.7	0.7	0.4
Precision@10 F	0.6	0.6	0.7	0.6	0.6	0.6
Avg Precision@10	0.54	0.54	0.58	0.60	0.58	0.54

Table 15: Summary of top-performing models' performance metrics when modeling *DC OBV* on the Wyscout dribbling dataset

For the Dribble and Carry (DC) *OBV* prediction on the Wyscout dataset, the MLP model slightly outperformed other models on MAE, while no linear model was in the top five performing models. The reduced linear model, which was the best performing linear model, was included for reference sake, but scored the 8th lowest MAE value. The MAE of the MLP was at a satisfactory level of 0.0216, considering the inter-quartile range of *DC OBV* was 0.042 (Figure 13). Random Forest models showed competitive performance, with various feature selection approaches yielding similar results. Precision@10 scores were high for fullbacks (FB) across most models, which implied that the Wyscout dataset contained features useful for identifying top-performing fullbacks based on their dribbling and carrying actions. For example, the MLP correctly predicted 8 of the top 10 *DC OBV* performing fullbacks in the test set.

6.3.1.2 Influence of features on model output

Figure 39 below shows the SHAP values for the top-performing MLP model on the dribbles and carries Wyscout dataset.

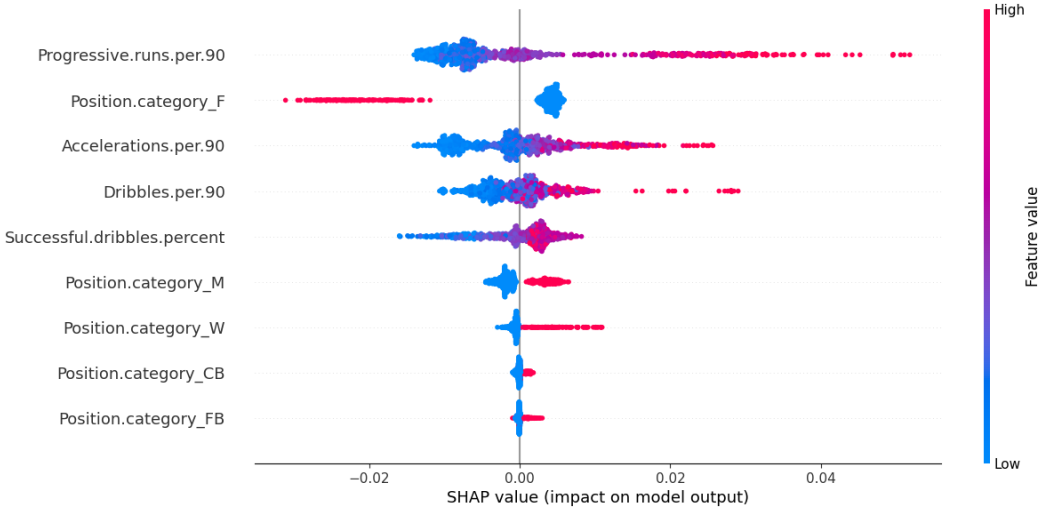

Figure 39: SHAP plot for MLP model fitted to Dribbles and carries Wyscout dataset

This SHAP plot for the MLP model on the DC Wyscout dataset illustrates that only a few variables held any importance in predicting *DC OBV*. *Progressive.runs.per.90* was the most important variable, with *Accelerations.per.90* and *Dribbles.per.90* also being important. Positioning variables showed that the MLP favoured wingers when it came to high *DC OBV* predictions, and forwards when it came to low *DC OBV* predictions.

6.3.1.3 Top-performing linear model

Table 16 below shows the reduced linear model on the dribbles and carries dataset. Playing as a winger (W) had the largest positive effect, followed by midfielder (M), fullback (FB), and center back (CB). *Progressive.runs.per.90* was a significant predictor of *DC OBV*. The presence of several significant interaction terms showed that the impact of dribbling actions varied by position. The Adjusted R-squared was 0.516, indicating that this model explained much less of the variance in *DC OBV* than the reduced linear model predicting *passing OBV*.

Variable	Beta	95% CI	p-value
(Intercept)	0.025	0.020, 0.030	≤ 0.001
W	0.041	0.032, 0.050	≤ 0.001
M	0.037	0.030, 0.044	≤ 0.001
CB	0.026	0.016, 0.036	≤ 0.001
FB	0.027	0.020, 0.034	≤ 0.001
Progressive runs per 90	0.027	0.021, 0.033	≤ 0.001
W * Progressive runs per 90	0.007	-0.002, 0.016	0.13
Accelerations per 90	0.007	0.001, 0.013	0.033
Successful dribbles percent	0.004	0.002, 0.007	≤ 0.001
FB * Accelerations per 90	-0.011	-0.019, -0.003	0.006
CB * Progressive runs per 90	-0.014	-0.024, -0.004	0.008
M * Progressive runs per 90	-0.014	-0.022, -0.006	≤ 0.001
W * Accelerations per 90	-0.015	-0.024, -0.005	0.002

Table 16: Regression results of reduced linear model fitted to dribbling actions Wyscout dataset, sorted by descending order of beta coefficients, with 95% Confidence Intervals (CI) and p-values

6.3.1.4 Top performing linear model: residual diagnostics

Figure 40 below displays the residual diagnostics of the reduced linear model fitted to the dribbles and carries Wyscout dataset.

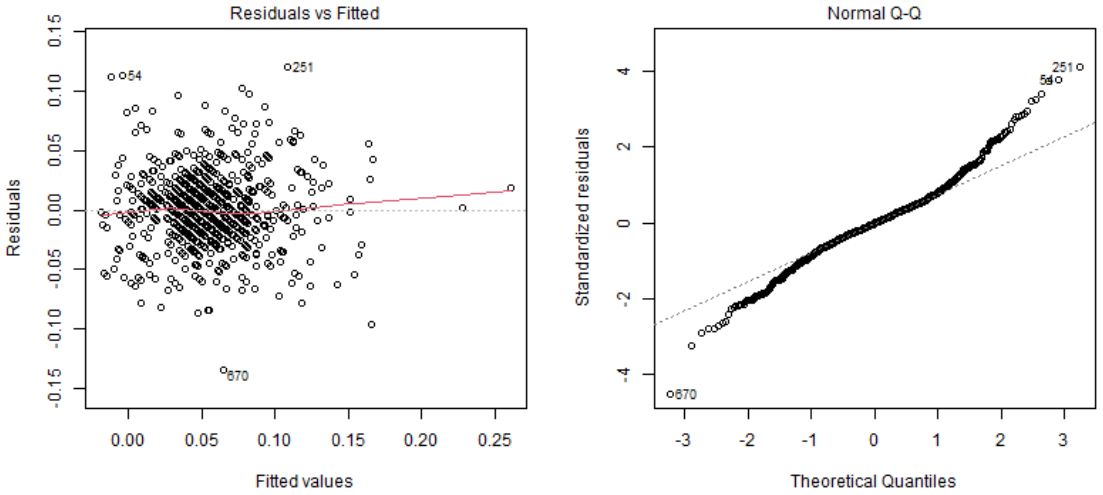

Figure 40: Linear Model on Defensive Wyscout dataset Diagnostic Plots

The Residuals vs Fitted plot reveals some heteroscedasticity, with variance increasing slightly for higher fitted values, but less pronounced than in the defensive action models. This showed more consistent predictive accuracy across different levels of *DC OBV*, however again not to the same level as the passing OBV models. A p-value of 2.2×10^{-16} on the Brauch-Pagan test confirmed the presence of heteroscedasticity. The Q-Q plot demonstrates that the residuals were once again not normally distributed, with heavy tails present, and a p-value of 2.23×10^{-10} on the Shaprio-Wilks test.

6.3.2 FBref Dataset

6.3.2.1 Top five models

Table 17 below shows the top five performing models fitted to the dribbles and carries FBref dataset.

Metric	RF Threshold	RF Original	RF Cumulative	RF Top N	Linear Model Reduced
Test Set MAE	0.0211	0.0213	0.0214	0.0214	0.0222
Test Set MSE	0.0009	0.0008	0.0009	0.0009	0.0010
Precision@5 CB	0.4	0.4	0.4	0.4	0.6
Precision@5 FB	0.8	0.8	0.8	0.8	0.8
Precision@5 M	0.8	0.8	0.8	0.8	0.8
Precision@5 W	0.4	0.4	0.6	0.4	0.6
Precision@5 F	0.8	0.8	0.8	0.8	0.8
Avg Precision@5	0.64	0.64	0.68	0.64	0.72
Precision@10 CB	0.8	0.8	0.8	0.8	0.9
Precision@10 FB	0.9	0.9	0.9	0.9	0.8
Precision@10 M	0.9	0.9	0.9	0.8	0.8
Precision@10 W	0.8	0.8	0.8	0.8	0.7
Precision@10 F	1.0	1.0	1.0	1.0	1.0
Avg Precision@10	0.92	0.92	0.92	0.90	0.88

Table 17: Summary of top-performing models' performance metrics modeling *DC OBV* on the FBref dribbling dataset

The Random Forest models consistently outperformed other model types(including linear models) for *DC OBV* prediction on the FBref dribbling dataset, with similar MAE and MSE scores across different Random Forest variations. Precision@10 scores were high across all positions, particularly for forwards (F) and fullbacks (FB). Precision@5 scores were high for all positions excluding wingers. Even though

the linear models didn't perform as well, the reduced model still achieved a significant Adjusted R^2 value of 0.711, and the highest Avg Precision@5 of the top 5 models. This suggested that the FBref dribbling and carrying features were highly informative for identifying top-performing players across most positions in terms of *DC OBV*.

6.3.2.2 Influence of features on model output

Figure 41 below shows the SHAP values for the top-performing Random Forest Threshold model on the FBref dribbles and carries dataset.

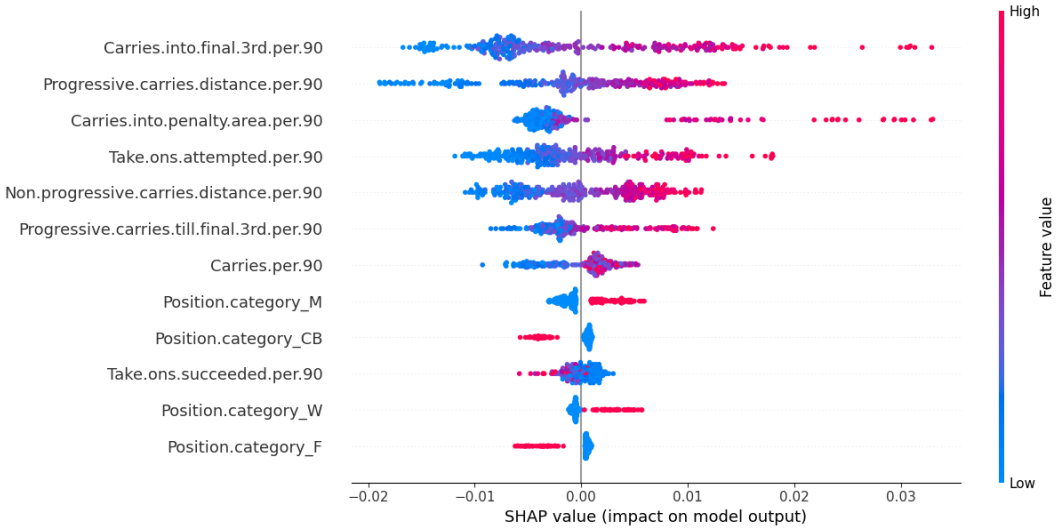

Figure 41: SHAP plot for Random Forest Threshold model fitted to dribbles and carries FBref dataset

Carries.into.final.3rd.per.90 and *Carries.into.penalty.area.per.90* were seen to be important metrics with high positive impact on the *DC OBV* predictions. The distance a player carries the ball also had a large impact on *DC OBV*, while position metrics favoured wingers and midfielders as these positions increased a player's *DC OBV* prediction.

6.3.2.3 Top-performing linear model

Table 18 below shows the reduced linear model fitted on the dribbles and carries FBref dataset.

Characteristic	Beta	95% CI	p-value
(Intercept)	0.071	0.065, 0.077	≤ 0.001
Progressive carries distance per 90	0.088	0.066, 0.110	≤ 0.001
Non-progressive carries distance per 90 * W	0.051	0.025, 0.077	≤ 0.001
M * Non-progressive carries distance per 90	0.046	0.022, 0.071	≤ 0.001
Non-progressive carries distance per 90 * CB	0.044	0.017, 0.071	0.001
Non-progressive carries distance per 90 * FB	0.040	0.010, 0.070	0.010
Take ons attempted per 90 * CB	0.023	0.011, 0.036	≤ 0.001
Carries into penalty area per 90	0.015	0.009, 0.020	≤ 0.001
M	0.014	0.005, 0.023	0.002
Take ons attempted per 90	0.005	-0.001, 0.011	0.10
Carries per 90	-0.006	-0.014, 0.002	0.20
Non-progressive carries distance per 90	-0.045	-0.069, -0.021	≤ 0.001
Progressive carries distance per 90 * W	-0.051	-0.076, -0.026	≤ 0.001
M * Progressive carries distance per 90	-0.059	-0.083, -0.035	≤ 0.001
Progressive carries distance per 90 * FB	-0.062	-0.091, -0.032	≤ 0.001
Progressive carries distance per 90 * CB	-0.070	-0.093, -0.046	≤ 0.001

Table 18: Regression results with 95% Confidence Interval (CI) and p-values of the model

Complex relationships between various carrying actions and *DC OBV* were evident. *Progressive.carries.distance.per.90* showed a strong positive effect, while *Non.progressive.carries.distance.per.90* had a significant negative effect on the predicted value. Interaction terms once again revealed that the impact of these actions varied significantly by position. An Adjusted R-squared of 0.677 indicated a moderately good fit.

6.3.2.4 Top performing linear model: residual diagnostics

Figure 42 below displays the residual diagnostics of the reduced linear model fitted to the dribbles and carries Wyscout dataset. While the residual plot seemingly reveals relatively constant variance across fitted values, a Breusch-Pagan test of heteroskedasticity yields a p-value of 2.19×10^6, indicating there were variables highly correlated with the variance of the residuals. The heavy tails of the Q-Q plot showed significant deviations from the assumed normal distribution, which was affirmed by a p-value of 2.38×10^6 obtained from the Shapiro-Wilks test .

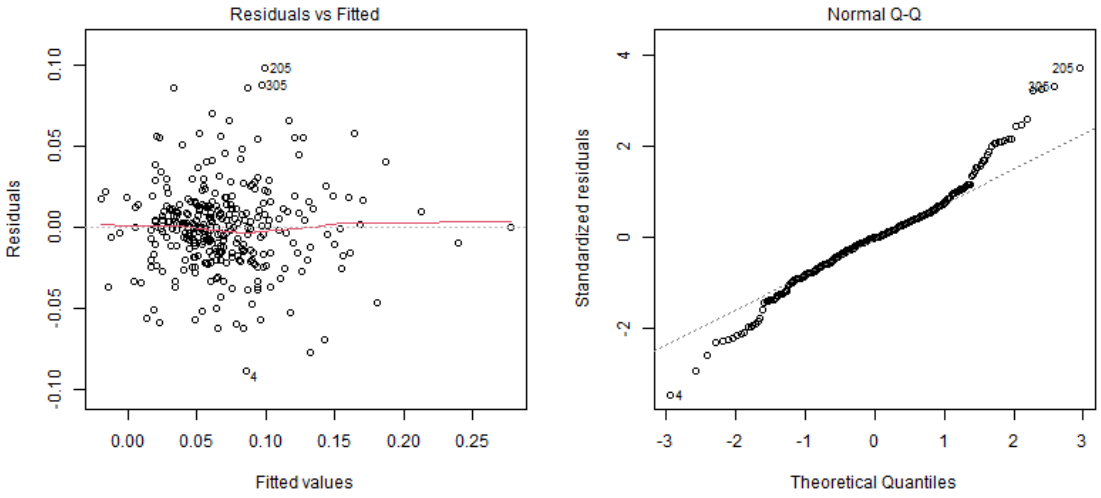

Figure 42: Linear Model on Defensive FBref dataset Diagnostic Plots

6.4 DA OBV

6.4.1 Wyscout Dataset

6.4.1.1 Top five models

Table 19 below shows the top 5 performing models fitted to the defensive actions Wyscout dataset. The Linear Models demonstrate the best overall performance with the lowest MAE and MSE scores, with the original model slightly outperforming the reduced one. However, the simpler linear model does have a lower AIC value (-3149.11) once again, when compared to the original linear model's AIC value(-3128.38). While these MAE and MSE values are similar to what was observed for the *Pass OBV*, the variance of the *DA OBV* is much smaller (about half), with most of the values existing within a range of 0.07. Precision@K values vary significantly across player positions, with fullbacks (FB) achieving the highest Precision@K score of 0.6. A result of note is that the top-performing linear model achieved a Precision@5 value of 0 for forwards, indicating of the top 5 predicted forwards, none of them belonged in the top 5. Interestingly, the MLP model shows competitive performance for Precision@K for forwards (F), indicating potential value for position-specific predictions.

Metric	LM Original	LM Reduced	MLP Original	XGB Cumulative	RF Original
Test Set MAE	0.0231	0.0232	0.0241	0.0252	0.0254
Test Set MSE	0.0011	0.0011	0.0013	0.0014	0.0014
Precision@5 CB	0.4	0.4	0.2	0.2	0.2
Precision@5 FB	0.6	0.4	0.2	0.0	0.0
Precision@5 M	0.4	0.4	0.2	0.2	0.2
Precision@5 W	0.2	0.2	0.2	0.0	0.2
Precision@5 F	0.0	0.2	0.6	0.6	0.4
Avg Precision@5	0.32	0.32	0.28	0.20	0.20
Precision@10 CB	0.5	0.4	0.4	0.4	0.4
Precision@10 FB	0.6	0.5	0.2	0.1	0.2
Precision@10 M	0.6	0.5	0.5	0.5	0.4
Precision@10 W	0.4	0.5	0.3	0.3	0.3
Precision@10 F	0.5	0.5	0.6	0.6	0.5
Avg Precision@10	0.52	0.48	0.40	0.38	0.36

Table 19: Summary of top-performing models' performance metrics when modeling *DA OBV* on the Wyscout defensive dataset

6.4.1.2 Influence of features on model output

Shap values for the best performing MLP model are shown in Figure 43 below.

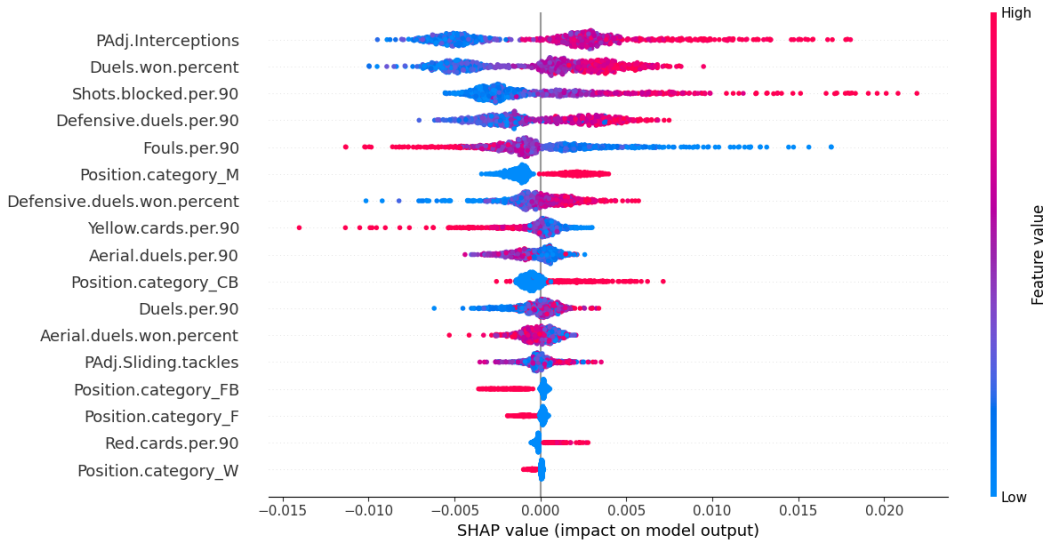

Figure 43: SHAP plot for MLP model fitted to Defnsive actions Wyscout dataset

PAdj.Interceptions(Possession adjusted interceptions, which are the number of interceptions a player makes multiplied by the percentage of time their team has the ball, which effectively weights a player's number of interceptions by the amount of opportunities there are to intercept the ball) and *Duels.won.percent* have the largest positive influence on the model's output, with high values for these features increasing the predicted value. In contrast, *Fouls.per.90* negatively affects the model. Some positions also have a significant effect, particularly midfielders and center-backs, while wingers have little effect on the model output.

6.4.1.3 Top-performing linear model

Table 20 below shows the best-performing linear model (reduced) on the defensive actions Wyscout dataset.

Characteristic	Beta	95% CI	p-value
(Intercept)	0.021	0.017, 0.026	≤ 0.001
M	0.018	0.011, 0.024	≤ 0.001
M * PAdj Interceptions	0.012	0.005, 0.018	≤ 0.001
Shots blocked per 90 * CB	0.007	0, 0.014	0.056
CB * PAdj Interceptions	0.007	0.001, 0.014	0.033
PAdj Interceptions * FB	0.006	-0.002, 0.014	0.12
Defensive duels per 90	0.006	0.003, 0.008	≤ 0.001
Shots blocked per 90	0.004	-0.001, 0.009	0.11
Yellow cards per 90	-0.003	-0.006, 0	0.029
Fouls per 90	-0.006	-0.009, -0.003	≤ 0.001

Table 20: Regression results for reduced linear model fitted to defensive actions Wyscout dataset, sorted by decreasing order of beta coefficients, with 95% Confidence Intervals (CI) and p-values

Playing as a midfielder had the largest positive effect, while the number of fouls a player commits had the largest negative impact on their predicted value. The presence of highly significant interaction terms, such as *PAdj.Interceptions* with the midfielder category, indicated that the importance of certain defensive actions varies by position. However, the very low Adjusted R-squared (0.174) implies that this model explains only a small proportion of the variance in *DA OBV* and that there are likely other significant factors not captured in this model.

6.4.1.4 Top performing linear model: residual diagnostics

Figure 44 below displays the residual diagnostics of the reduced linear model fitted to the defensive actions Wyscout dataset.

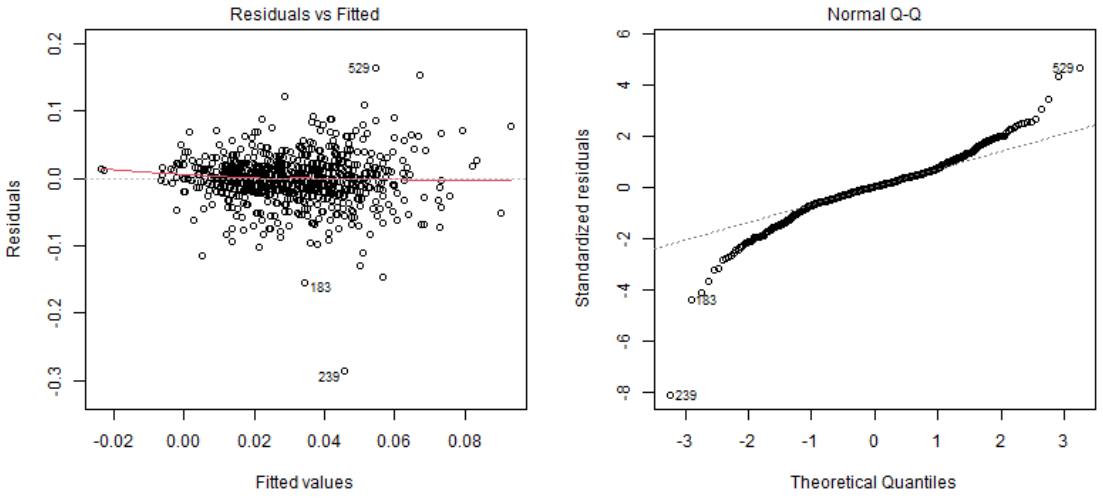

Figure 44: Residual diagnostics for reduced linear model fitted to defensive actions Wyscout dataset

The Residuals vs Fitted plot revealed some heteroscedasticity, with variance increasing for higher fitted values, indicating that the model's predictive accuracy varies across different levels of *DA OBV*. The presence of heteroscedasticity is confirmed by a Breusch- Pagan p-value of 5.27×10^{27}. The Q-Q plot demonstrated that the tails of the residuals were far too heavy and so the assumption of normal residuals was violated. This was confirmed by a Shapiro-Wilks test p-value of 2.2×10^{16}.

6.4.2 FBref Dataset

6.4.2.1 Top five models

Table 21 below shows the top five performing models fitted to the defensive actions FBref dataset. Similar test set MAE values to the Wyscout DA dataset were observed. The MLP model outperformed the other models (0.0271 test set MAE). Precision@K scores were generally higher across all positions when compared to the Wyscout dataset, particularly for midfielders (M) and wingers (W). This implied that the FBref dataset contained more informative features for defensive actions, especially for non-defender positions. Linear models again under performed, with the reduced model being the sole linear model in the top 5 by MAE.

Metric	MLP Original	LM Original	RF PCA 99%	XGB PCA 80%	LM Reduced
Test Set MAE	0.0271	0.0293	0.0293	0.0294	0.0295
Test Set MSE	0.0015	0.0016	0.0018	0.0018	0.0017
Precision@5 CB	0.2	0.4	0.4	0.4	0.4
Precision@5 FB	0.4	0.6	0.8	0.8	0.8
Precision@5 M	0.8	0.8	0.4	0.4	0.4
Precision@5 W	0.6	0.6	0.0	0.4	0.4
Precision@5 F	0.4	0.4	0.2	0.6	0.6
Avg Precision@5	0.48	0.56	0.36	0.52	0.52
Precision@10 CB	0.5	0.5	0.6	0.5	0.5
Precision@10 FB	0.8	0.8	0.9	0.9	0.9
Precision@10 M	0.8	0.8	0.7	0.6	0.6
Precision@10 W	0.9	0.9	0.6	0.7	0.7
Precision@10 F	0.7	0.6	1.0	1.0	1.0
Avg Precision@10	0.74	0.72	0.76	0.74	0.74

Table 21: Summary of top-performing models' performance metrics modeling *DA OBV* on the FBref defensive dataset

6.4.2.2 Influence of features on model output

Figure 45 below shows the SHAP values for the top-performing MLP model on the FBref defensive actions dataset. *Interceptions.per.90* had the largest positive impact on the model's output, a result that was found in the corresponding Wyscout dataset too. *Tackles.in.defensive.-3rd.per.90* and *Tackles.in.midfield.3rd.per.90* also contribute positively, though to a lesser extent. Player positions were once again important, particularly forwards, which had a negative effect on model output.

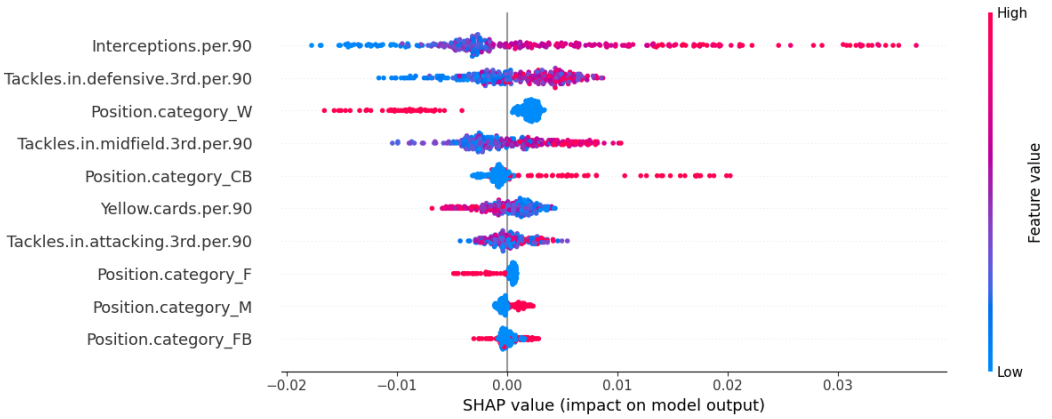

Figure 45: SHAP plot for MLP model fitted to defensive actions FBref dataset

6.4.2.3 Top-performing linear model

Table 22 below shows the reduced linear model on the defensive actions FBref dataset.

Characteristic	**Beta**	**95% CI**	**p-value**
(Intercept)	0.032	0.027, 0.037	≤ 0.001
Interceptions per 90 * CB	0.023	0.012, 0.034	≤ 0.001
Interceptions per 90	0.012	0.006, 0.018	≤ 0.001
Tackles in attacking 3rd per 90	0.004	-0.001, 0.009	0.092
Tackles in midfield 3rd per 90	0.004	-0.001, 0.009	0.094
Yellow cards per 90	-0.006	-0.010, -0.001	0.021
W * Interceptions per 90	-0.016	-0.036, 0.004	0.11
W	-0.023	-0.042, -0.004	0.016

Table 22: Regression results of reduced linear model fitted to defensive actions FBref dataset, sorted by descending order of beta coefficients, with 95% Confidence Intervals (CI) and p-values

Notably, playing as a winger (W) has a negative effect, while *Interceptions.per.90* is strongly positively associated with *DA OBV*. The interaction term of *Interceptions.per.90* and *CB* had a large positive coefficient, highlighting the particular importance of interceptions for center backs. The Adjusted R-squared (0.207) was found to be higher than that for the Wyscout model, but was still very low.

6.4.2.4 Top performing linear model: residual diagnostics

Figure 46 below displays the residual diagnostics of the reduced linear model fitted to the defensive actions Wyscout dataset.

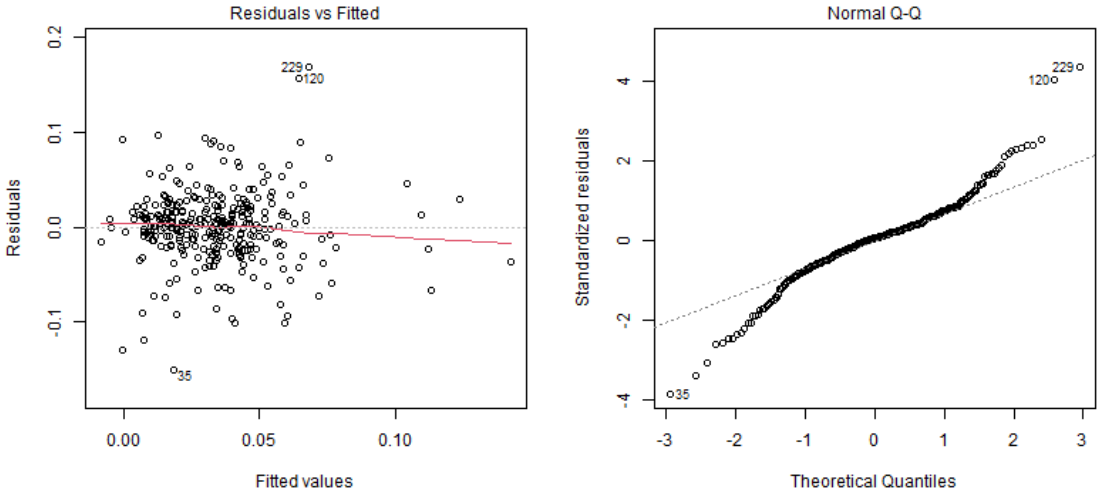

Figure 46: Residual diagnostics for reduced linear model fitted to defensive actions FBref dataset

The Residuals vs Fitted plot revealed some heteroscedasticity, with variance increasing for higher fitted values. This was backed up by a Breusch-Pagan p-value of 0.004. The significant deviations from the norm in the Q-Q plot also demonstrated further deviation from the linear model's assumptions, with the tails seemingly very heavy. This was illustrated by a Shapiro-Wilks test p-value of 4.99×10^{-9}.

6.5 Decision Support System (DSS)

The DSS developed is designed to predict *OBV* and its components using Wyscout data, and thus is completely applicable for PSL clubs. It features four main tabs: the Home tab provides a landing page, the Process Files tab allows users to upload Wyscout data, the Visualize Data tab displays the resulting predictions, and the Budget Scouting tab offers an enhanced visualization that integrates budget considerations into the scouting process. The DSS is thus an all-encompassing scouting tool, allowing for raw insights to be turned into budget adjusted scouting profiles, per *OBV* component.

The application landing page is a simple page with a tab-bar menu on the left to navigate through the app (Figure 47 below).

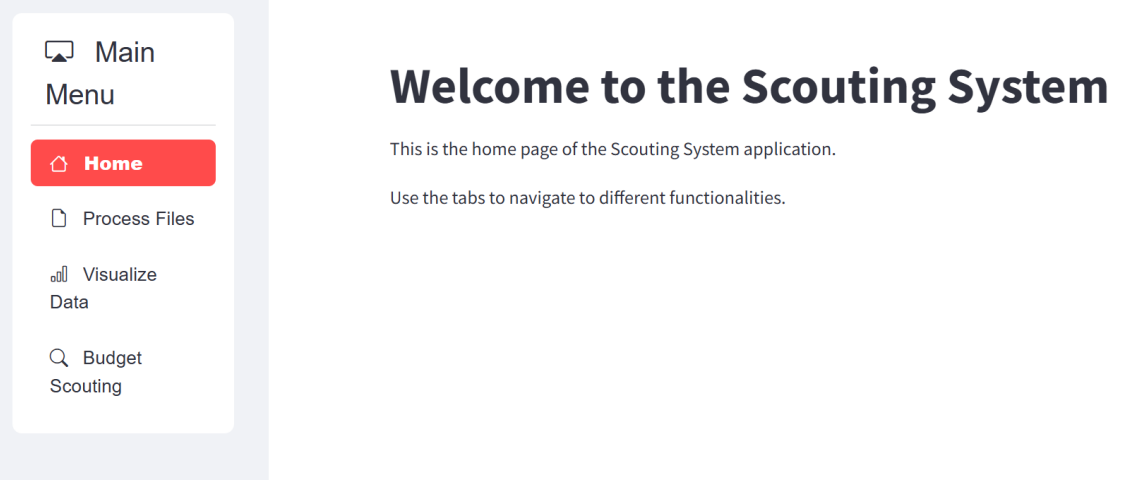

Figure 47: Landing page of DSS

The `Process files` page is where the Wyscout player data can be uploaded into the app. This page allows files to be dragged-and-dropped or alternatively manually uploaded by clicking the button Browse Files (see Figure 48 below). The name of the file is automatically set to `output.csv` but can be manually changed depending on the input file. Once `Process Files` is clicked, all the data is cleaned, the relevant models are initialized and *OBV* predictions are made. The xlsx file with the final predictions is also generated and saved.

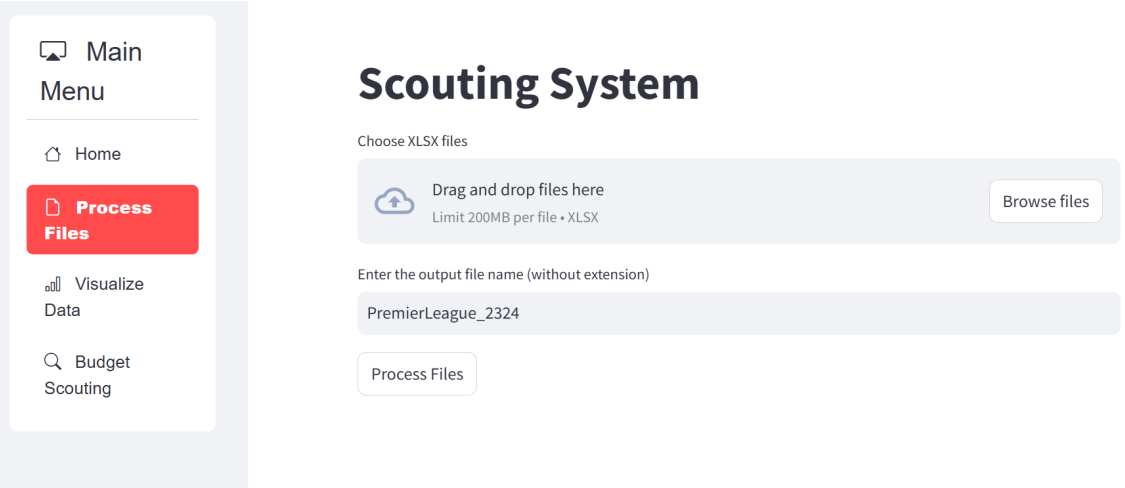

Figure 48: Data upload page of DSS

Upon moving to the `Data Visualization` page (Figure 49 below), the user has the option to choose which file they want to analyse. Each season of a specific league is a separate file. In Figure 49 the last season's Premier League, in England, has been chosen. This selection will load in the OBV predictions made for all players in this league for the specified season. The specific position of interest must then be selected. After that, the OBV components the user wants to analyse must be selected. One can select any OBV component (or *Total* OBV), and a maximum of two selections at a time can be made.

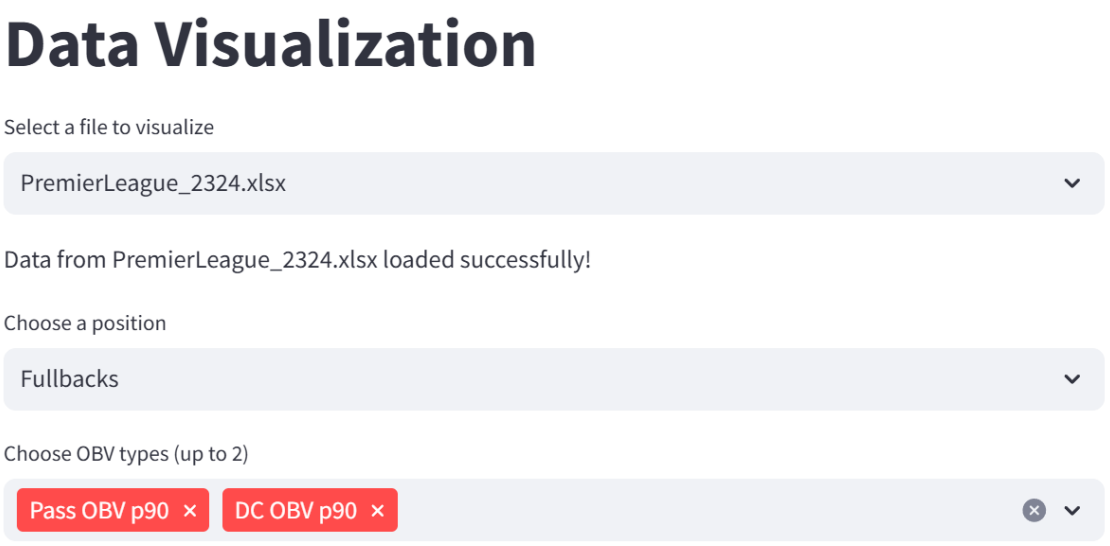

Figure 49: Inputs on `Data Visualization tab`

If a single *OBV* type is selected, then a barchart is displayed (Figure 50 below). Since the package plotly (Inc., 2015) is used, the charts have interactive capabilities, such as when hovering over the bars, one can force the player's name and *OBV* values to be shown. The data downloaded from Wyscout for this dissertation, however, does not contain player names and so it is not shown below.

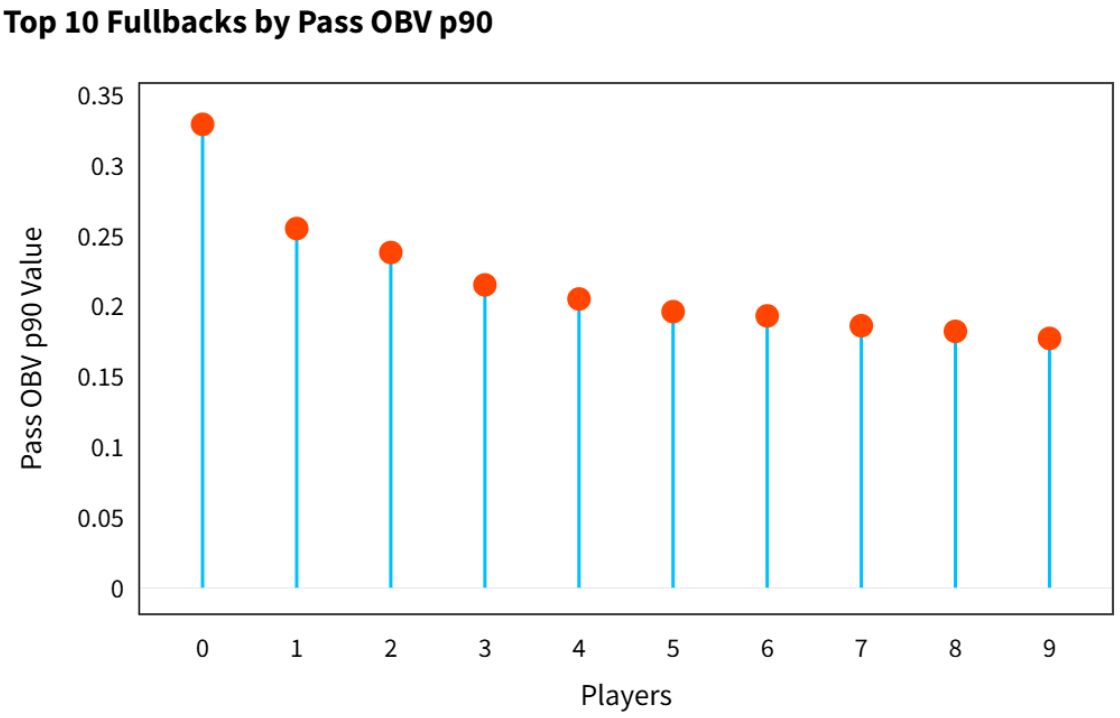

Figure 50: Barchart on `Data Visualization` page

If two *OBV* types are selected, then a scatter plot is shown, such as the one in Figure 51 below. The order of the axes follows the order of the components selected in Figure 50 above, with the first component plotted as the x-axis and the second component as the y-axis.

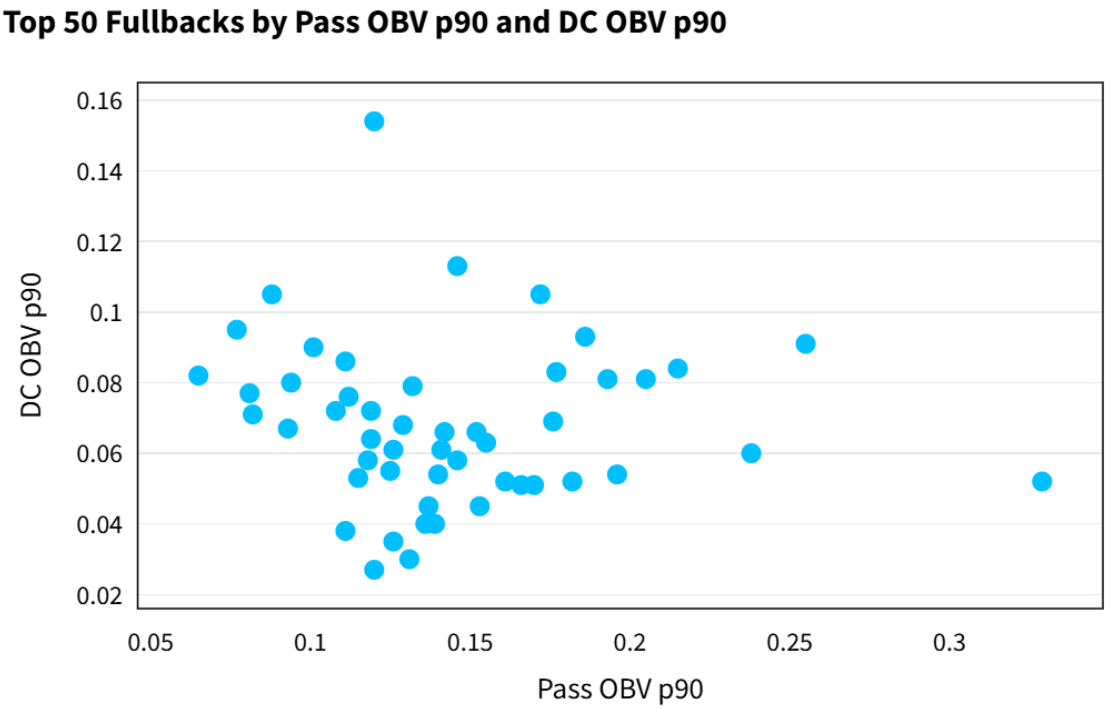

Figure 51: Scatter plot on `Data Visualization` page

The `Budget Scouting` tab (Figure 52) allows a user to incorporate their budget restrictions into their player scouting search. The slider in Figure 52 below demonstrates the user's ability to restrict the player search to only include players who cost between a certain upper and lower bound.

Figure 52: Inputs on the `Budget Scouting` page

The cost per OBV value attained can also be observed, through charts like Figure 53 below. The bar chart illustrates the top 10 fullbacks ranked by the scaled result of dividing OBV by the cost of the player. This plot thus displays the value-for-money performance between the players as opposed to solely the quality of their performance.

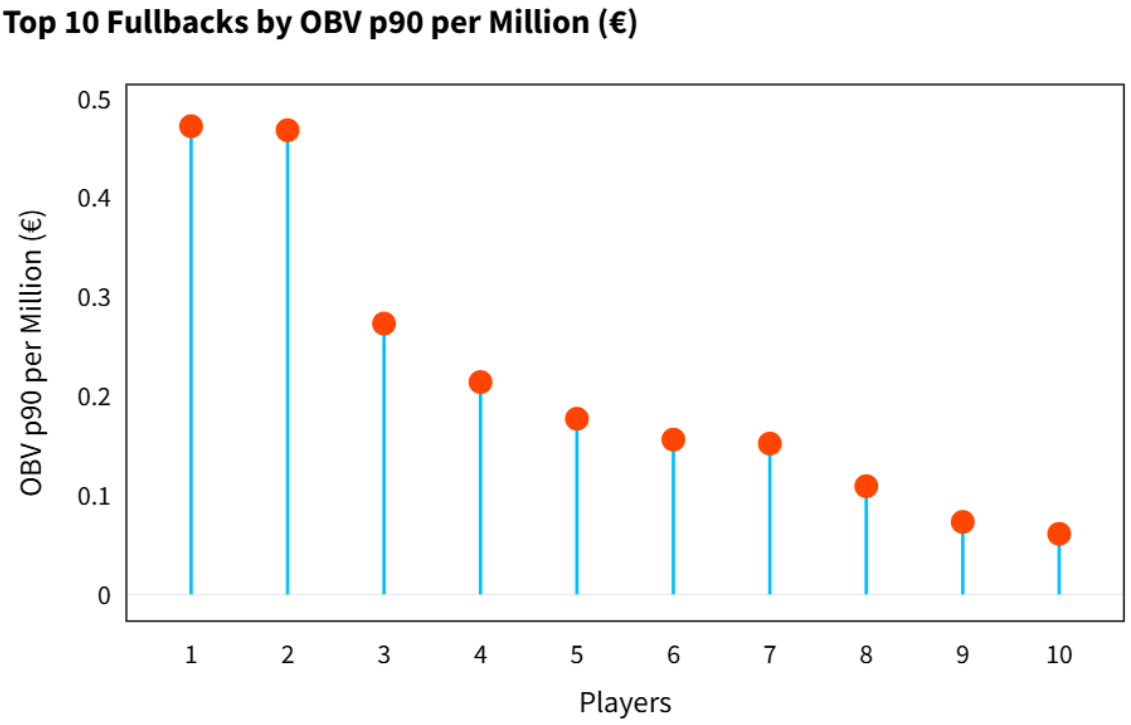

Figure 53: Barchart showing *OBV* per million euros

A final scatter plot in Figure 54 below illustrates the relationship between players' market value (x-axis) and their *OBV* p90 performance (y-axis), with red dotted lines dividing the plot into four distinct quadrants. The upper-left quadrant shows players with above-average *OBV* but below-average market value and the lower-left quadrant displays players with below average *OBV* and below average market value. The lower-right quadrant shows players with below-average *OBV* but above-average market value and the upper right quadrant players with above average *OBV* and market value.

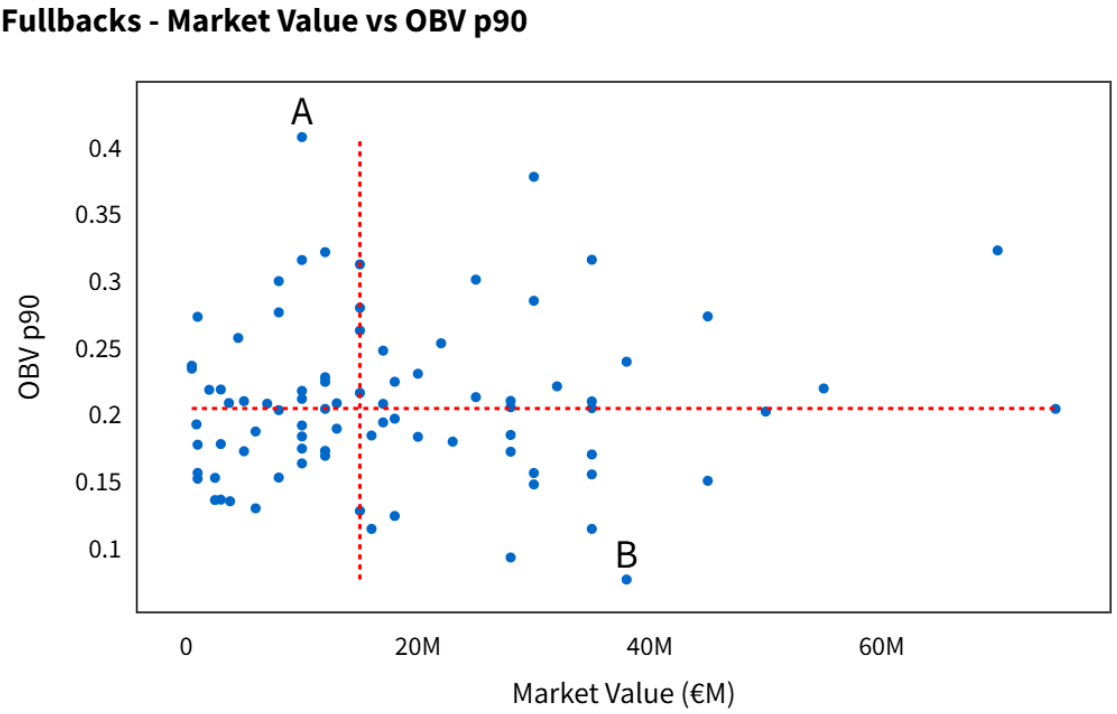

Figure 54: Player *OBV* vs player market Value

6.6 Conclusion

This section successfully demonstrates the feasibility and effectiveness of modeling On-Ball Value (*OBV*) and its components using readily available event-frequency data, revealing a clear performance difference across different methodologies. While simpler, more interpretable linear models consistently provided the most accurate predictions for metrics *Total OBV* and *Pass OBV*, Random Forests and MLPs had to be employed to achieve optimal performance for *DC OBV* and *DA OBV*. Across all models and datasets, a consistent narrative identified progressive actions—particularly progressive passes and carries—as the most significant drivers of on-ball value, with SHAP analyses and interaction terms confirming that the value of these actions is critically dependent on player position. Although the top-performing linear models exhibited strong predictive utility, diagnostic assessments revealed significant heteroscedasticity and non-normal residuals, indicating their statistical inferences should be interpreted with caution. The culmination of this work is a practical DSS that operationalizes these findings, translating complex model outputs into an interactive platform that integrates budget constraints, thereby providing clubs with a robust, low-cost framework for identifying and evaluating talent through data-driven insights.

7 Discussion

7.1 Modeling Discussion

The analysis of *Total OBV* using the Wyscout dataset revealed that linear regression models achieved the best performance compared to the more complex ML techniques, a result that may stem from the data lacking nonlinear, complex interactions. The reduced linear model incorporated features which were previously hypothesised to be important (Table 2 in the EDA). *Progressive.passes.per.90* was found to be highly significant (p-value$\leq$ 0.01), and had the 3rd largest coefficient value (0.032), indicating an additional progressive pass per game increased the predicted *Total OBV* for a player by 0.032. These positive results were expected, as Table 2 revealed moderate correlations between several explanatory variables and *Total OBV*. Although residual diagnostics indicated some heteroscedasticity, the overall reliability of the linear model remained high enough to warrant its practical utility. The reduced linear model's ability to correctly identify seven of the top ten wingers in the test set was particularly noteworthy, suggesting that scouts could use this approach to predict a significant portion of the top performers across more than 50 leagues worldwide. This level of scope is especially valuable when contrasted with the high costs associated with buying more extensive data from providers such as StatsBomb. Moreover, the fact that this modeling used only frequency data—much cheaper and simpler to collect—yet managed to decently capture much of the complexity of *OBV* highlights the viability and importance of this approach. Successfully relying on frequency metrics constitutes a significant step toward the democratization of player evaluation, as this data can be more readily acquired by smaller clubs or individual analysts who do not have large budgets or advanced technical resources.

The subsequent modeling of *OBV* using the FBref dataset further corroborated the effectiveness of linear regression models in capturing *Total OBV*. In this case, the model achieved an improved adjusted R^2 of 0.734, compared to 0.709 when the Wyscout dataset was used, and demonstrated improved Precision@k scores. These results were consistent with what was previously hypothesized in the EDA, as the correlations between variables in the FBref dataset and *Total OBV* were generally larger than between variables in the Wyscout dataset and *Total OBV*. Furthermore, the location-informed variables in the FBref dataset proved a significant differentiator between the linear model trained on the Wyscout dataset and the linear model trained on the FBref dataset. Variables such as *Progressive.passing.distance.per.90* and *Progressive.carries.distance.per.90* yielded significant impact on the prediction of *Total OBV*, with interactions between these variables and the player's position, as informed by Figure 21 and Figure 24, creating a further in-depth understanding of where the actions were likely taking place. The impressive results of the linear

model mean that, for example, analysts can immediately apply the model to any of the top five European leagues—where FBref frequency data is available—to extract the top 10 best wingers by *OBV*, at no cost. Such a result is highly advantageous for teams aiming to adopt data-driven scouting methods without incurring large data acquisition costs, thus lowering the barrier to entry for clubs looking to integrate analytics into their recruitment strategies.

When *Pass OBV* was modeled using Wyscout data, linear regression models again produced the lowest test set MAE. The results of the linear model included Precision@10 scores for defenders and midfielders that indicated that 6 of the top 10 players in the test set were correctly identified. This outcome suggests that a simple, easily implemented linear approach can estimate player passing ability moderately well. An adjusted R^2 of 0.758 further attested to the model's ability to explain a substantial portion of the variance in passing performance, which was an improvement on the linear model trained on the full dataset when predicting *Total OBV* (0.709). The better performance over complex ML models is again likely due to the lack of complex interactions in the data, while numerous moderate correlations were present, allowing the linear model to perform well. Interestingly, of the five variables in the Wyscout passing dataset that were most correlated with *Pass OBV*, only *Progressive.passes.per.90* was in the final reduced linear model. Other variables moderately correlated with *Pass OBV* -*Passes.to.final.third.per.90* and *Passes.to.penalty.area.per.90*- were both excluded in the final model, while the number of crosses from the right and left flank were seen as highly influential with coefficients of 0.09 and 0.08, respectively, and p-values$\leq$0.001. However, the detection of heteroskedasticity and non-normal residuals indicates some instability, suggesting that these results should ideally be corroborated with qualitative analyses such as video scouting. Nevertheless, these results highlight that frequency statistics can be used to accurately predict *Pass OBV*, a value usually derived from complex and expensive event-based data. This implies that PSL teams and analysts can generate meaningful insights about a players passing quality without relying on costly, high-resolution tracking data. This not only reduces the financial burden associated with player evaluation but also broadens the accessibility of where teams can scout, enabling clubs to look abroad with confidence in their data-driven decision-making.

Linear models were again the strongest predictors of *Pass OBV*, this time when modeled using FBref data. The linear models significantly out-performed the ML models, with a drop of roughly 32% in test set MAE when comparing the closest performing ML model (RF Top N) with the reduced linear model. These results were unsurprising as Table 3 showed that the passing dataset from FBref showed some of the highest correlations with *OBV* (or an *OBV* component). Variables such as *Long.passes.per.90*, *Crosses.into.penalty.area* and *Progressive.passes.per.90*

all exhibited correlations with *Pass OBV* of more than 0.5, suggesting that linear models would adequately capture the variance in *Pass OBV*. As all these variables achieved p-values $\leq$0.001 and the optimal linear model reached an average Precision@10 score of 0.84 across player positions, and an improved adjusted R^2 of 0.829, these prior intuitions were vindicated. The performance of the linear model using free FBref data demonstrates that an enhanced understanding of a player's passing ability is achievable at no cost for analysts, making advanced player evaluation more accessible.

While linear models demonstrated strong performance in modeling Passing and *Total OBV*, this success did not extend to *DC OBV*, especially when training on the Wyscout dataset, where linear relationships were far less evident. The correlation analysis in on Table 4 indicated that models fit to the Wyscout dribbling dataset were likely going to struggle to explain the variation in *DC OBV*, since only two variables had correlation values greater than 0.5 with *DC OBV*. Furthermore, as previously hypothesized and seen, variables with location information embedded in them are crucial to predicting *OBV*, however *Progressive.runs.per.90* (highest correlation with *DC OBV*) was the only variable in the data set of this type. In light of this, ML techniques demonstrated superior performance, with the MLP model slightly outperforming numerous Random Forest models. The optimal MLP model achieved an average Precision@10 score of 0.54. This result implied that approximately half of the players identified among the top 10 across all positions were correctly selected. The reduced linear model achieved the same Precision@10 score, and had an adjusted R^2 value of 0.516, which illustrated that the model left a significant amount of the variation of *DC OBV* unexplained. The SHAP plot for the MLP further highlighted why models struggled to capture the variance of *DC OBV*, with an over-reliance on variables *Progressive.runs.per.90*, *Accelerations.per.90* and *Dribbles.per.90*, with the latter two variables not accounting for any spatial information. From this it was evident that the utility of the model was limited, as its predictive power remained constrained by the lack of spatial data in the available features. As a result, while the model can provide a useful preliminary screening tool, as shown by its moderate Precision@10 score of 0.54, its predictions should be interpreted with caution and supplemented with video scouting (accessible through Wyscout at no extra cost) to validate the findings of the model.

When modeling *DC OBV* with the FBref dataset, the results were more favourable than when *DC OBV* was modeled using the Wyscout dataset. In this instance, linear models again performed worse than their ML counterparts, while Random Forest models proved particularly effective. The location-enriched data was shown to be significantly correlated with *DC OBV*, with the top five variables from Table 4 all being location-informed variables, and all having correlations greater than 0.5

with *DC OBV*. This led to the reduced linear model attaining an improved adjusted R^2 value of 0.677, showcasing the positive effect the locational-data had on the model. The SHAP plot further emphasized the importance of the spatial aspect of frequency statistics, with five of the six most influential variables containing spatial components, and all ranking above player position in terms of influence. The optimal Random Forest model achieved an average Precision@10 value of 0.92, indicating a high proficiency in accurately identifying the top 10 dribblers. Building upon the earlier findings that the FBref passing dataset can effectively be used to model *Pass OBV*, these results indicate that a Random Forest model trained on the FBref dribbling dataset offers a robust, cost-effective method for identifying top dribblers from the five leading European leagues.

When modeling *DA OBV* with the Wyscout dataset, the results were unsatisfactory. Although linear regression models exhibited the best test set error metrics, a more detailed examination revealed that four of the five top-performing models yielded Precision@10 scores below 0.5. Moreover, the reduced linear model achieved an adjusted R^2 of only 0.174, indicating that a substantial portion of the variance in *DA OBV* remained unexplained. These outcomes were consistent with what was presupposed during numerous preliminary analyses in the EDA. Firstly, Figure 7 showed that *DA OBV* was closely linked to location, and thus any explanatory variables that would sufficiently capture the variance of *DA OBV* would have to be correspondingly granular. Upon examining the variables in Table A2 and the correlation table (Table 5), it became evident that the available variables were too broad in their spatial representation, with correlations too weak to effectively explain *DA OBV* (the highest correlation being only 0.333). This assertaion was proved correct in the results with the optimal linear model achieving an adjusted R^2 of only 0.174. These findings confirm that more granular spatial data is needed to capture the intricate variance in *DA OBV*. Future data collection efforts could thus be informed by the insights gleaned from Figure 7, where new defensive variables could be engineered to specifically target the high-value defensive zones, thereby enhancing the ability to model *DA OBV* in the future.

When modeling *DA OBV* using the FBref dataset, a similar challenge was expected and observed. Although the pitch was divided into thirds to capture more detailed tackle location data, the level of spatial resolution remained too coarse to accurately predict defensive action *OBV*. This was illustrated by Figure 7, which showed considerable variability of *DA OBV*, even within key areas such as the penalty box. While the FBref dataset provided variables like *Tackles.in.defensive.3rd.per.90*, *Tackles.in.midfield.3rd.per.90*, and *Tackles.in.attacking.3rd.per.90*, these broad categorizations did not adequately capture the fine-grained locational context needed to account for *DA OBV*. The reduced linear model underscored this limitation by

dropping *Tackles.in.defensive.3rd.per.90* entirely, along with the MLP model finding *Interceptions.per.90* to be more important than *Tackles.in.defensive.3rd*, highlighting its weak explanatory power. Nonetheless, under these constraints, the MLP model achieved an average Precision@10 of 0.72, indicating that although the absolute predictive scores remain constrained by the lack of granularity in the data, the relative ordering of players may still provide meaningful insights into their defensive adequacy. Consequently, analysts could use an MLP model as a preliminary screening tool, generating shortlists of promising defensive performers before refining their assessments with more detailed video-based evaluations.

7.2 DSS Discussion

The decision support system (DSS) presented in this study demonstrates a feasible implementation of an easy-to-use scouting prototype that integrates predictive modeling of OBV and related components with budget-based decision-making. The system illustrates the ease with which the process of uploading, cleaning, and modeling Wyscout data can be undertaken. This shows how efficiently clubs in the PSL can seamlessly move from raw datasets to actionable insights. By blending data visualization and modeling with budget-aware scouting, the DSS provides multiple layers of analysis that could aid decision-makers in identifying both high-impact performers and cost-effective transfer targets.

The DSS comprises tools designed to fit directly into existing club workflows. The results from the model building phase would be readily available for inspection within the DSS interface, where large volumes of players can be filtered and ranked with minimal user input. By offering simple drag-and-drop file processing and interactive visualizations, the system ensures that analysts do not require extensive technical expertise to move from raw data to actionable scouting insights. The ability to switch between single and multiple OBV metrics provides further user customization options, illustrating the system's flexibility to explore focused performance dimensions or to compare two types of contributions simultaneously.

Budgetary factors are also incorporated into the DSS through a dedicated tab that enables the specification of upper and lower bounds on market value. The cost-per-OBV metric can be used to highlight players who deliver strong performances relative to their market value, ensuring PSL clubs can easily identify high-performing, yet undervalued talent. Not only does this combined emphasis on budget and performance strengthen a club's ability to make well-informed transfer decisions, it also facilitates the quick filtering of large candidate player pools.

8 Conclusion

This dissertation has presented a series of linear and ML models capable of estimating *OBV* and its components using both Wyscout and FBref datasets. Across both datasets, linear regression emerged as a surprisingly robust predictor for *Total* and *Pass OBV*, accurately identifying a substantial fraction of the top performers in the test sets. These findings indicate that the more parsimonious linear framework can capture key performance metrics, which rely heavily on location, with the given frequency data. Additional modeling of *DC OBV* and *DA OBV* underscored the lack of information in these datasets, as both linear and ML approaches often lacked significant explanatory power to fully account for variability in these metrics. However, in instances where predictive accuracy was moderate, the models still produced respectable precision@10 scores, indicating that their generated candidate lists could still offer value for scouts who may then choose to supplement these findings with other analyses.

The DSS prototype also shows how these models can be effectively integrated into a user-friendly workflow. By offering straightforward file processing, interactive data visualization, and budget-aware scouting features, the DSS could potentially reduce the technical burden on club staff and enable efficient filtering and ranking of large player pools. Incorporating cost-per-*OBV* metrics further extends the scope of utility as one can evaluate both performance impact and economic feasibility, which is particularly advantageous for clubs facing financial constraints, such as many of those in the PSL. Although initially built on Wyscout data, the DSS can easily accomodate FBref data, enabling usage in clubs that wish to scout across only Europe's top five leagues. Consequently, a range of clubs, from the PSL to those in Europe's top five competitions, can readily adopt and benefit from this scouting system.

8.1 Limitations and future research

Although the techniques presented here offer a promising foundation for data-driven scouting in a budget-constrained landscape, several limitations do exist. Firstly, a key limitation of the current *OBV* estimates is that they treat all leagues as equal in strength, which is unlikely to be true. The next phase of research could thus focus on incorporating league strengths into the calculations. By developing league-specific strength estimates and integrating them into the predictions, one could better contextualize player valuations based on their competition level, ultimately leading to more accurate assessments.

A further constraint involves the datasets themselves. The FBref component of this

project only corresponded to one season of StatsBomb data, limiting both model calibration and the generalizability of the results obtained from models fitted to this data. Incorporating additional StatsBomb data, alongside expanded Wyscout and FBref explanatory datasets, would provide a richer empirical foundation and potentially more robust estimates. In additiona to this, more detailed spatial data for defensive actions would greatly enhance the accuracy of *DA OBV* models. Another methodological improvement that could increase the predictive capability of the ML models would be the introduction of interaction terms between frequency-based variables and the player's position in the datasets, since they proved highly significant in the linear models. This omission could have prevented the ML models from fully leveraging the relationships captured by these interactions. Moreover, future research could be improved by grouping positions differently. For example, midfielders could be categorised as attacking and defensive midfielders. Employing more fine-grained positional definitions could improve estimates by allowing interaction terms to reflect more precisely the specific areas of the pitch where different progressive actions are likely taking place.

Future research into *OBV* prediction using frequency data could strengthen the linear models by applying different transformations to address nonlinear relationships and uneven error variance—such as employing logarithmic transformations or power transformations on explanatory and *OBV* variables. Weighted least squares could also be used to correct for heteroscedasticity, leading to more accurate coefficient estimates and better model generalization. for linear models.

Despite these constraints, the core findings of this dissertation demonstrate the cost-effectiveness of linear models for various *OBV* metrics and the practicality of the DSS for rapid data processing and budget-informed scouting within both the PSL and Europe.

References

Aalbers, B., & Haaren, J. V. (2018). Distinguishing between roles of football players in play-by-play match event data [Preprint]. https://www.researchgate.net/publication/328148296

Abreu, P. H., Silva, D. C., Portela, J., & Reis, L. P. (2014). Using model-based collaborative filtering techniques to recommend the expected best strategy to defeat a simulated soccer opponent [Accessed: 2024-04-01]. *Intelligent Data Analysis*, *18*, 973–991. https://doi.org/10.3233/IDA-140678

Ahsan, M., Mahmud, M., Saha, P., Guo, Y., & Siddiquee, M. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52. https://doi.org/10.3390/technologies9030052

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alcaraz-Herrera, H., Cartlidge, J., Toumpakari, Z., Western, M., & Palomares, I. (2022). Evorecsys: Evolutionary framework for health and well-being recommender systems [Accessed: 2024-04-01]. *User Modeling and User-Adapted Interaction*, *32*, 883–921. https://doi.org/10.1007/s11257-022-09320-3

Aygün, M., Savaş, Y., & Savaş, D. A. (2023). The relation between football clubs and economic growth: The case of developed countries [Published: 13 September 2023]. *Humanities and Social Sciences Communications*, *10*, Article number: 566. https://doi.org/10.1057/s41599-023-02074-2

Azmat, A., & Yi, S. S. (2024). Machine learning [arXiv:2401.08718 [cs.LG]]. *arXiv preprint*. https://doi.org/10.48550/arXiv.2401.08718

Bargagli Stoffi, F. J., Cevolani, G., & Gnecco, G. (2022). Simple models in complex worlds: Occam's razor and statistical learning theory. *Minds and Machines*, *32*(1), 13–42. https://doi.org/10.1007/s11023-022-09592-z

Beal, R., Norman, T. J., & Ramchurn, S. D. (2019). Artificial intelligence for team sports: A survey [Accessed: 2024-04-01]. *The Knowledge Engineering Review*, *34*. https://doi.org/10.1017/S0269888919000225

Beernaerts, J., Baets, B. D., Lenoir, M., & de Weghe, N. V. (2022). Qualitative team formation analysis in football: A case study of the 2018 fifa world cup [Accessed: 2024-04-01]. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.863216

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, I-115–I-123. http://proceedings.mlr.press/v28/bergstra13.pdf

Berrar, D., Lopes, P., & Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine Learning*, *108*(7). https://doi.org/10.1007/s10994-018-5747-8

Bertheussen, B. A. (2023). Hard talk, costly walk: The evolution of a soft budget constraint syndrome in a football club at the periphery of europe. *Frontiers in Sports and Active Living*, *5*, 1107988. https://doi.org/10.3389/fspor.2023.1107988

Bransen, L., Haaren, J. V., & van de Velden, M. (2019). Measuring soccer players' contributions to chance creation by valuing their passes. *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2018-0020

Breiman, L. (2001a). Random forests [Accessed: 2024-04-01]. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294. https://doi.org/10.2307/1911963

Brooks, J., Kerr, M., & Guttag, J. (2016). Developing a data-driven player ranking in soccer using predictive model weights. *Proceedings of the 22nd ACM SIGKDD International Conference*. https://doi.org/10.1145/2939672.2939695

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *3*(2), 1–30. https://doi.org/10.1145/3105576

Chen, Q., & Qi, J. (2023). How much should we trust r^2 and adjusted r^2: Evidence from regressions in top economics journals and monte carlo simulations. *Journal of Applied Economics*, *26*(1), 2207326. https://doi.org/10.1080/15140326.2023.2207326

Chen, T., & Guestrin, C. (2016a). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. https://doi.org/10.1145/2939672.2939785

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

Decroos, T., Bransen, L., Haaren, J. V., & Davis, J. (2018). Actions speak louder than goals: Valuing player actions in soccer [Significant update of the paper with a clearer methodology, applied on a different data set, and more extensive experiments. Published at SIGKDD 2019]. *arXiv preprint arXiv:1802.07127*. https://doi.org/10.48550/arXiv.1802.07127

Developers, S. (2024). Streamlit: The fastest way to build and share data apps [Accessed: April 2024]. https://streamlit.io/

Dick, U., & Brefeld, U. (2022). Action rate models for predicting actions in soccer. *AStA Advances in Statistical Analysis*, *107*(1). https://doi.org/10.1007/s10182-022-00435-x

Duch, J., Waitzman, J. S., & Amaral, L. A. N. (2010). Quantifying the performance of individual players in a team activity [Accessed: 2024-04-01.]. *PLoS ONE*, *5*(6), e10937. https://doi.org/10.1371/journal.pone.0010937

El-Roby, A., Hefny, A., & Choubineh, A. (2023). *Stratalign: Uncovering tactical patterns through large-scale event sequence matching* [Accessed: 2024-04-01].

https://statsbomb.com/wp-content/uploads/2023/10/Ahmed_El-Roby-2.pdf

FBref. (2025). Fbref: Advanced football stats and history [Accessed: 2024-03-28]. https://fbref.com/en/

Felfernig, A., Wundara, M., Tran, T. N. T., Le, V.-M., et al. (2024). *Recommender systems for sustainability: Overview and research issues* [Accessed: 2024-04-01]. https://doi.org/10.48550/arXiv.2412.03620

FIFA. (n.d.). Stadium guidelines [Accessed: 2024-04-19]. https://inside.fifa.com/technical/stadium-guidelines

Forcher, L., Altmann, S., Forcher, L., Jekauc, D., & Kempe, M. (2022). The use of player tracking data to analyze defensive play in professional soccer - a scoping review. *International Journal of Sports Science & Coaching*, *17*(6). https://doi.org/10.1177/1747954122107573

Freedman, D. A. (2009). *Statistical models: Theory and practice* (2nd). Cambridge University Press.

Garratt-Stanley, F. (2023a, May). *Inside the rise of roberto de zerbi's brighton & hove albion* [Accessed: 2024-04-19]. https://www.jobsinfootball.com/blog/inside-the-rise-of-roberto-de-zerbis-brighton-hove-albion

Garratt-Stanley, F. (2023b, May). *Inside the rise of roberto de zerbi's brighton & hove albion* [Accessed: April 2024]. https://www.jobsinfootball.com/inside-the-rise-of-roberto-de-zerbis-brighton-hove-albion

Green, S. (2012). Assessing the performance of premier league goalscorers [Accessed April 2024]. https://www.statsperform.com

Gyarmati, L., & Stanojevic, R. (2016). Qpass: A merit-based evaluation of soccer passes [2016 ACM KDD Workshop on Large-Scale Sports Analytics]. *arXiv preprint arXiv:1608.03532*. https://doi.org/10.48550/arXiv.1608.03532

Han, S., & Kim, H. (2021). Optimal feature set size in random forest regression. *Applied Sciences*, *11*(8), 3428. https://doi.org/10.3390/app11083428

Herberger, T. A., & Litke, C. (2021a). The impact of big data and sports analytics on professional football: A systematic literature review [Accessed: 2024-04-01.]. In T. A. Herberger & J. J. Dötsch (Eds.), *Digitalization, digital transformation and sustainability in the global economy* (pp. 147–171). Springer. https://doi.org/10.1007/978-3-030-77340-3_12

Herberger, T. A., & Litke, C. (2021b). The impact of big data and sports analytics on professional football: A systematic literature review [Handle: RePEc:spr:prbchp:978-3-030-77340-3$_1$2]. In T. A. Herberger & J. J. Dötsch (Eds.), *Digitalization, digital transformation and sustainability in the global economy* (pp. 147–171). Springer. https://doi.org/10.1007/978-3-030-77340-3_12

Hinton, G., & Roweis, S. (2002). Stochastic neighbor embedding. *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'02)*, 857–864. https://dl.acm.org/doi/10.5555/2968618.2968725

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hudl StatsBomb. (2021a). *Introducing on-ball value (obv)* [Accessed: 2024-03-28]. https://statsbomb.com/news/introducing-on-ball-value-obv/

Hudl StatsBomb. (2021b, July). Learn more about the hudl statsbomb aggregated api [Accessed: April 2024]. https://statsbomb.com/articles/soccer/learn-more-about-the-statsbomb-iq-api/

Hudl StatsBomb. (2024). Statsbomb: Advanced football analytics [Accessed: 2024-03-28]. https://statsbomb.com/product-login/

Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports [Published Online: August 21, 2019]. *International Journal of Computer Science in Sport*, *18*(1), 1–23. https://doi.org/10.2478/ijcss-2019-0001

Hvattum, L. M., & Sæbø, O. D. (2015). Evaluating the efficiency of the association football transfer market using regression-based player ratings. *NIK: Norsk Informatikkonferanse*. https://www.researchgate.net/publication/283320847_Evaluating_the_efficiency_of_the_association_football_transfer_market_using_regression_based_player_ratings

Inc., P. T. (2015). *Collaborative data science*. https://plot.ly

Kharrat, T., McHale, I. G., & Peña, J. L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, *283*, 726–736. https://doi.org/10.1016/j.ejor.2019.11.052

King, W., & Bhorat, T. (2022). *Using the counterfactual to value defensive actions in women's football* [Honours thesis]. University of Cape Town.

Klemp, M., Wunderlich, F., & Memmert, D. (2021). In-play forecasting in football using event and positional data. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-03157-3

Lawlor, C., Rookwood, J., & Wright, C. (2021). Player scouting and recruitment in english men's professional football: Opportunities for research. *Journal of Qualitative Research in Sports Studies*, *15*(1), 57–76. https://www.academia.edu/60947249/Craig_Lawlor_Joel_Rookwood_and_Craig_Wright_2021_Player_scouting_and_recruitment_in_English_men_s_professional_football_opportunities_for_research_Journal_of_Qualitative_Research_in_Sports_Studies_15_1_57_76

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine learning and conventional statistics: Making sense of the differences. *Knee*

Surgery, Sports Traumatology, Arthroscopy, *30*, 753–757. https://doi.org/10.1007/s00167-022-06896-6

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0168768

Lolli, L., Bauer, P., Irving, C., Bonanno, D., Höner, O., Gregson, W., & Salvo, V. D. (2024). Data analytics in the football industry: A survey investigating operational frameworks and practices in professional clubs and national federations from around the world [Accessed: 2025-01-28. Licensed under Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0.]. *Science and Medicine in Football*, *n/a*, 1–10. https://doi.org/10.1080/24733938.2024.2341837

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions [Version 2, last revised 25 Nov 2017]. *arXiv preprint arXiv:1705.07874*. https://arxiv.org/abs/1705.07874

Matano, F., Richardson, L. F., Pospisil, T., Eubanks, C., & Qin, J. (2018). Augmenting adjusted plus-minus in soccer with fifa ratings [Submitted on 18 Oct 2018]. *arXiv preprint arXiv:1810.08032*. https://doi.org/10.48550/arXiv.1810.08032

McDonnell, B., & Sisneros, M. (2023, March). *Brighton and brentford: Two smart clubs who play the game in opposite ways* [Accessed: 2024-04-19]. https://theanalyst.com/2023/03/brighton-and-brentford-two-smart-clubs-who-play-the-game-in-opposite-ways

McHale, I. G., & Relton, S. D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, *268*(1), 339–347. https://doi.org/10.1016/j.ejor.2018.01.025

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). On the development of a soccer player performance rating system for the english premier league. *Interfaces*, *42*(4). https://doi.org/10.1287/inte.1110.0589

Merhej, C., Ryan, J. B., Matthews, T., & Ramchurn, S. D. (2021). What happened next? using deep learning to value defensive actions in football event-data [Presented at KDD 2021, Virtual Conference.]. *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2021)*. https://doi.org/10.1145/3447548.3467090

Mirkowicz, M., & Grodner, G. (2018). Jakob nielsen's heuristics in selected elements of interface design of selected blogs. *Social Communication*, *4*(2), 30–51. https://doi.org/10.2478/sc-2018-0013

Modric, T., Versic, S., & Sekulic, D. (2020). Aerobic fitness and game performance indicators in professional football players; playing position specifics and associations. *Heliyon*, *6*(11), e05427. https://doi.org/10.1016/j.heliyon.2020.e05427

Morrow, S. (2023, January). Football's ever-changing economics. In *The people's game?* (pp. 1–51). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-031-12917-5_1

Naylor, A., et al. (2023, April). *How brighton's transfers have become the envy of the premier league* [Accessed: 2024-04-19]. https://www.nytimes.com/athletic/4237821/2023/04/21/brighton-scouting-premier-league-tony-bloom/

O'Brien, S. (2020, January). *Doing it differently how brentford flipped the script and staged a data revolution to become england's smartest club* [Accessed: 2024-04-19]. https://talksport.com/football/659667/brentford-data-revolution-england-smartest-club-championship-leicester-fa-cup/#:~:text=%22It's%20not%20that%20data%20tells,higher%20level%20than%20people%20think.

octosport.io. (2022). Expected goals: Can they predict future goals? [Published in *Geek Culture*. Accessed April 2024]. https://medium.com/geekculture/expected-goals-can-they-predict-future-goals-24a1b1d0279d

Opta. (2025). Opta player stats by statsperform: Soccer statistics and analytics [Accessed: 2024-03-28]. https://optaplayerstats.statsperform.com/en_GB/soccer

Otte, F., Dittmer, T., & West, J. (2022). Goalkeeping in modern football: Current positional demands and research insights. *International Sport Coaching Journal*, *10*(1), 112–120. https://doi.org/10.1123/iscj.2022-0012

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Peterson, K., & Evans, L. (2019). Decision support system for mitigating athletic injuries [Accessed: 2024-04-01]. *International Journal of Computer Science in Sport*, *18*(1), 45–63. https://doi.org/10.2478/ijcss-2019-0003

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1605–1613. https://doi.org/10.1145/3097983.3098051

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest? *Journal of Machine Learning Research*, *18*(1), 6673–6690. https://jmlr.org/papers/volume18/17-269/17-269.pdf

Putatunda, S., & Rama, K. (2020). A modified bayesian optimization based hyperparameter tuning approach for extreme gradient boosting. *arXiv preprint arXiv:2004.05041*. https://arxiv.org/abs/2004.05041

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rajesh, V., Arjun, P., Jagtap, K. R., M., S. C., & Prakash, J. (2022). Player recommendation system for fantasy premier league using machine learning [Accessed: 2024-04-01]. *Proceedings of the 2022 19th International Joint Conference*. https://ieeexplore.ieee.org/document/9836260

Rohde, M., & Breuer, C. (2016). Europe's elite football: Financial growth, sporting success, transfer investment, and private majority investors [Accessed: 2024-04-01.]. *International Journal of Financial Studies*, *4*(2), 1–20. https://doi.org/10.3390/ijfs4020012

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sæbø, O. D., & Hvattum, L. M. (2018). Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics*, *4*, 55–66. https://doi.org/10.3233/JSA-170235

Santos-Gago, J. M., Ramos-Merino, M., Vallarades-Rodriguez, S., Álvarez-Sabucedo, L. M., Fernández-Iglesias, M. J., & García-Soidán, J. L. (2019). Innovative use of wrist-worn wearable devices in the sports domain: A systematic review [Accessed: 2024-04-01]. *Electronics*, *8*(11), 1257. https://doi.org/10.3390/electronics8111257

Schultze, S. R., & Wellbrock, C.-M. (2017). A weighted plus/minus metric for individual soccer player performance [First published online: October 12, 2017]. *Journal of Sports Analytics*. https://doi.org/10.3233/JSA-170225

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. https://www.jstor.org/stable/2333709

Silva, V. D., Caine, M., Skinner, J., Dogan, S., Kondoz, A., Peter, T., Axtell, E., Birnie, M., & Smith, B. (2018). Player tracking data analytics as a tool for physical performance management in football: A case study from chelsea football club academy. *Sports*, *6*(4), 130. https://doi.org/10.3390/sports6040130

Singh, K. (2020). Introducing expected threat (xt): Modelling team behaviour in possession to gain a deeper understanding of buildup play [Accessed April 2024]. https://karun.in/blog/expected-threat.html

SkillCorner. (2024). Skillcorner: Ai-powered football tracking [Accessed: 2024-03-28]. https://skillcorner.com/

Stats Perform. (2024). Introducing a possession value framework [Accessed April 2024]. https://www.statsperform.com/resource/introducing-a-possession-value-framework/

StatsBomb. (2021). *What is expected threat (xt)? possession value models explained* [Accessed: 2024-04-01]. https://statsbomb.com/soccer-metrics/possession-value-models-explained/

Sumpter, D. (2016, May). *Soccermatics: Mathematical adventures in the beautiful game pro-edition*. Bloomsbury Sigma.

Szymanski, S. (2001). Income inequality, competitive balance and the attractiveness of team sports: Some evidence and a natural experiment from english soccer. *The Economic Journal*, *111*(469), F69–F84. https://www.jstor.org/stable/2667958

Thakkar, P., & Shah, M. (2021). An assessment of football through the lens of data science. *Annals of Data Science*, *8*, 823–836. https://doi.org/10.1007/s40745-021-00323-2

Transfermarkt. (n.d.). Betway premiership [Accessed: 2024-04-19]. https://www.transfermarkt.co.za/betway-premiership/startseite/wettbewerb/SFA1

Triady, M. S., & Utami, A. F. (2015). Analysis of decision making process in moneyball: The art of winning an unfair game. *The Winners*, *16*(1), 57. https://doi.org/10.21512/tw.v16i1.1555

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne [Accessed: 2024-04-01]. *Journal of Machine Learning Research*, *9*, 2579–2605. https://www.researchgate.net/publication/228339739_Visualizing_data_using_t-SNE

Van Roy, M., Robberechts, P., Yang, W.-C., De Raedt, L., & Davis, J. (2021). Learning a markov model for evaluating soccer decision making [Presented at Reinforcement Learning for Real Life Workshop, July 23, 2021, Virtual. Accepted version. Open Access.]. *Reinforcement Learning for Real Life (RL4RealLife) Workshop at ICML 2021*. https://lirias.kuleuven.be/handle/123456789/3498953

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, *11*(44). https://doi.org/10.1186/s40537-024-00858-w

Warnke, A. J., & Sittl, R. (2016). *Competitive balance and assortative matching in the german bundesliga* (Working Paper) (Version: 16-058, State: In Progress,

Published Online: August 2016). ZEW Discussion Paper. https://doi.org/10.13140/RG.2.2.15203.55843

Watanabe, S. (2023). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*. https://arxiv.org/abs/2304.11127

Whitmore, J. (2020). Evolving our possession value framework [Accessed April 2024]. https://www.statsperform.com/resource/evolving-our-possession-value-framework/

Wigmore, T. (2017, July). *Brentford's moneyball way to beat football teams with huge budgets* [Accessed: 2024-04-01]. https://bleacherreport.com/articles/2718752-brentfords-moneyball-way-to-beat-football-teams-with-huge-budgets

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis [Accessed: 2024-04-01]. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wolf, S., Schmitt, M., & Schuller, B. (2020). A football player rating system [Licensed under CC BY-NC 4.0]. *Journal of Sports Analytics*, *6*(5), 1–15. https://doi.org/10.3233/JSA-200411

Wu, F., Wang, Q., Bian, J., Xiong, H., Ding, N., Lu, F., Cheng, J., & Dou, D. (2022). A survey on video action recognition in sports: Datasets, methods and applications [Accessed: 2024-04-01]. *arXiv preprint*. https://doi.org/10.48550/arXiv.2206.01038

Wu, M., Kolen, J., Aghdaie, N., & Zaman, K. A. (2017). Recommendation applications and systems at electronic arts [Accessed: 2024-04-01]. *RecSys '17: Proceedings of the Eleventh ACM Conference on Recommender Systems*, 338. https://doi.org/10.1145/3109859.3109928

Wyscout. (2024). Wyscout api documentation: Comprehensive football data access [Accessed: 2025-04-28]. https://apidocs.wyscout.com/

Yiapanas, G., Thrassou, A., & Vrontis, D. (2024). The contemporary football industry: A value-based analysis of social, business structural and organisational stakeholders [Accessed: 2024-04-19]. *Accounting, Auditing & Accountability Journal*, *37*(2), 552–585. https://doi.org/10.1108/AAAJ-06-2022-5855

Yılmaz, Ö. İ., & Öğüdücü, Ş. G. (2022). Learning football player features using graph embeddings for player recommendation system [Accessed: 2024-04-01]. *SAC '22: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 577–584. https://doi.org/10.1145/3477314.3507257

Zhang, W., Gong, B., Tao, R., Zhou, F., Ruano, M. Á. G., & Zhou, C. (2024). The influence of tactical formation on physical and technical performance across playing positions in the chinese super league. *Scientific Reports*, *14*, 2538. https://doi.org/10.1038/s41598-024-53113-0

A Appendix

A.1 Statsbomb data

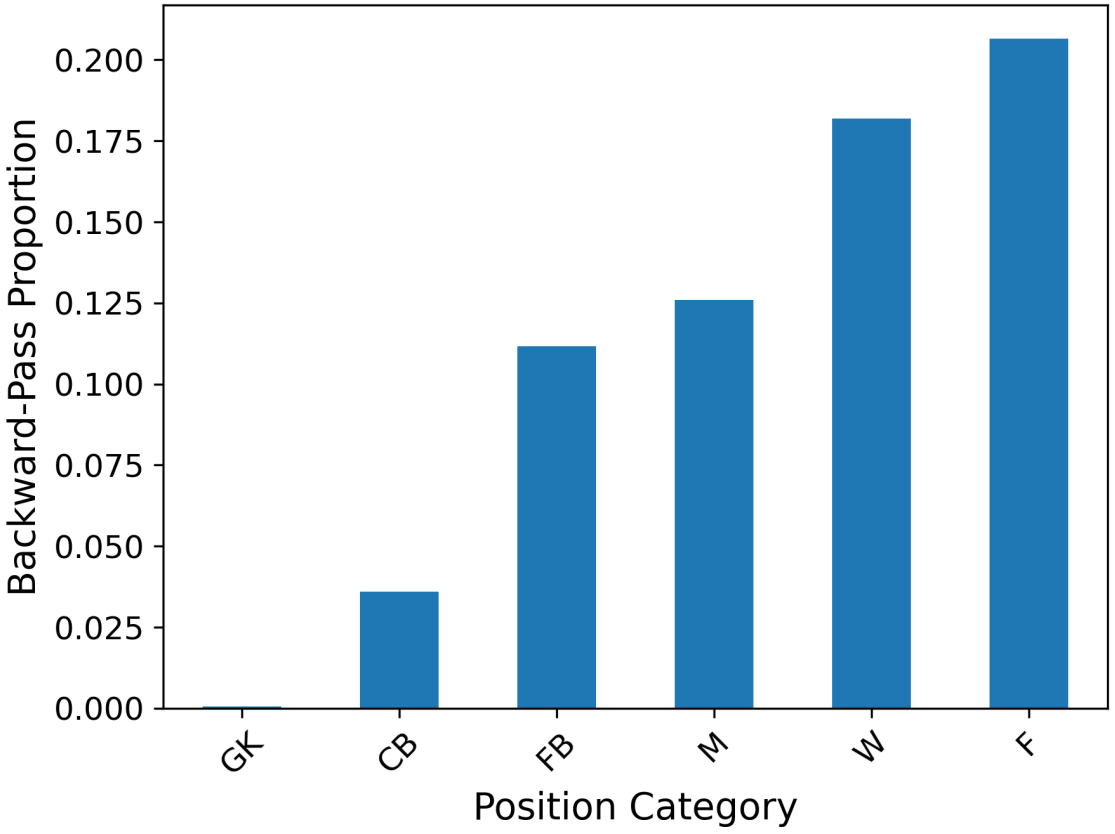

Figure A1: Proportion of passes played backward by player position

A.2 Wyscout Variables

Variable	Description
Accurate.crosses.percent	Percentage of accurate crosses
Crosses.from.left.flank.per.90	Crosses from the left flank per 90 minutes
Accurate.crosses.from.left.flank.percent	Percentage of accurate crosses from the left flank
Crosses.from.right.flank.per.90	Crosses from the right flank per 90 minutes
Accurate.crosses.from.right.flank.percent	Percentage of accurate crosses from the right flank
Crosses.to.goalie.box.per.90	Crosses to the goalie box per 90 minutes
Back.passes.per.90	Back passes per 90 minutes
Accurate.back.passes.percent	Percentage of accurate back passes
Short.medium.passes.per.90	Short and medium passes per 90 minutes
Accurate.short.medium.passes.percent	Percentage of accurate short and medium passes
Average.pass.length.m	Average pass length in meters
Accurate.long.pass.length.m	Accurate long pass length in meters
xA.per.90	Expected assists per 90 minutes
Smart.passes.per.90	Smart passes per 90 minutes
Accurate.smart.passes.percent	Percentage of accurate smart passes
Passes.to.final.third.per.90	Passes to the final third per 90 minutes
Accurate.passes.to.final.third.percent	Percentage of accurate passes to the final third
Passes.to.penalty.area.per.90	Passes to the penalty area per 90 minutes
Accurate.passes.to.penalty.area.percent	Percentage of accurate passes to the penalty area
Progressive.passes.per.90	Progressive passes per 90 minutes
Accurate.progressive.passes.percent	Percentage of accurate progressive passes

Table A1: Wyscout Passing Variables

Variable	**Description**
Duels.per.90	Duels per 90 minutes
Duels.won.percent	Percentage of duels won
Defensive.duels.per.90	Defensive duels per 90 minutes
Defensive.duels.won.percent	Percentage of defensive duels won
Aerial.duels.per.90	Aerial duels per 90 minutes
Aerial.duels.won.percent	Percentage of aerial duels won
PAdj.Sliding.tackles	Possession-adjusted sliding tackles
Shots.blocked.per.90	Shots blocked per 90 minutes
PAdj.Interceptions	Possession-adjusted interceptions
Fouls.per.90	Fouls per 90 minutes
Yellow.cards.per.90	Yellow cards per 90 minutes
Red.cards.per.90	Red cards per 90 minutes

Table A2: Wyscout Defensive Variables

Variable	**Description**
Dribbles.per.90	Dribbles per 90 minutes
Successful.dribbles.percent	Percentage of successful dribbles
Progressive.runs.per.90	Progressive runs per 90 minutes
Accelerations.per.90	Accelerations per 90 minutes

Table A3: Wyscout Dribbling Variables

A.3 FBref Variables

Variable	Description
Progressive.passes.till.final.3rd.per.90	Progressive passes into the final third per 90 minutes
Non.progressive.passing.distance.per.90	Non-progressive passing distance per 90 minutes
Assists.per.90	Assists per 90 minutes
Assists.Minus.Expected.Assisted.goals.per.90	Assists minus expected assisted goals per 90 minutes
Expected.assists.per.90	Expected assists per 90 minutes
Long.passes.Per.90	Long passes per 90 minutes
Accurate.long.passes.percent	Percentage of accurate long passes
Accurate.medium.passes.percent	Percentage of accurate medium passes
Short.passes.per.90	Short passes per 90 minutes
Accurate.short.passes.percent	Percentage of accurate short passes
Progressive.passing.distance.per.90	Progressive passing distance per 90 minutes
Passes.into.final.3rd.per.90	Passes into the final third per 90 minutes
Passes.into.penalty.area.per.90	Passes into the penalty area per 90 minutes
Crosses.into.penalty.area	Crosses into the penalty area
Through.balls.per.90	Through balls per 90 minutes
Progressive.passes.received.per.90	Progressive passes received per 90 minutes

Table A4: FBref Passing Variables

Variable	Description
Yellow.cards.per.90	Yellow cards per 90 minutes
Red.cards.per.90	Red cards per 90 minutes
Tackles.in.attacking.3rd.per.90	Tackles in the attacking third per 90 minutes
Tackles.in.defensive.3rd.per.90	Tackles in the defensive third per 90 minutes
Tackles.in.midfield.3rd.per.90	Tackles in the midfield third per 90 minutes
Interceptions.per.90	Interceptions per 90 minutes

Table A5: FBref Defensive Variables

Variable	Description
Progressive.carries.till.final.3rd.per.90	Progressive carries into the final third per 90 minutes
Progressive.carries.distance.per.90	Progressive carries distance per 90 minutes
Non.progressive.carries.distance.per.90	Non-progressive carries distance per 90 minutes
Take.ons.attempted.per.90	Take-ons attempted per 90 minutes
Take.ons.succeeded.per.90	Take-ons succeeded per 90 minutes
Carries.into.final.3rd.per.90	Carries into the final third per 90 minutes
Carries.into.penalty.area.per.90	Carries into the penalty area per 90 minutes
Carries.per.90	Carries per 90 minutes

Table A6: FBref Dribbling Variables

A.4 Correlation Analysis

A.4.1 Wyscout Dataset

The variables listed in Tables A1,A2,and A3 above, provide a comprehensive overview of a player's average (per 90 minutes) performance across various facets of the game. These variables can be grouped into categories of passing (Table A1), dribbling (Table A2), and defending (Table A3), which are the 3 OBV components that will be modelled. Passing metrics are extensive, detailing the player's distribution accuracy and effectiveness. Key variables include *Passes.per.90* and *Accurate.passes.percent*, which measure the frequency and accuracy of passes. Specific passing scenarios are also captured, such as *Forward.passes.per.90, Accurate.forward.passes.percent, Back.passes.per.90*, and *Accurate.back.passes.percent*. Advanced metrics like *Smart.passes.per.90* and *Key.passes.per.90* reflect the player's playmaking abilities, while *Progressive.passes.per.90* and *Accurate.progressive.passes.percent* indicate their success in advancing the ball forward. Additionally, crossing ability is detailed with variables like *Accurate.crosses.percent* and *Crosses.to.goalie.box.per.90*, that record the precision and effectiveness in delivering crosses.

Dribbling involves the player's ability to manoeuvre the ball and bypass opponents. Important variables include *Dribbles.per.90* and *Successful.dribbles.percent*, which quantify the frequency and success rate of dribbles. Variables such as *Progressive.runs.per.90* and *Accelerations.per.90* measure their capability to carry the ball forward and make advancements up the field through carrying the ball.

Defending encompasses metrics that evaluate a player's defensive prowess. Variables such as *Duels.per.90, Duels.won.percent, Defensive.duels.per.90*, and *Defensive.duels.won.percent* assess their engagement and success in defensive battles. *Successful.defensive.actions.per.90* captures overall defensive effectiveness, while *Interceptions.per.90* and *PAdj.Interceptions.per.90* focus on their ability to intercept passes. Additional defensive actions are highlighted by *Sliding.tackles.per.90, Shots.blocked.per.90*, and *Fouls.per.90*, which provide insight into tackling, shot-blocking, and disciplinary record. Notably, the absence of locational defensive metrics presents a significant challenge, as it is likely to hinder the development of a robust model for *DA OBV*.

Figure A2 below shows the correlation between Wyscout passing variables. Several key insights emerge from this analysis. First, pairs of variables such as *Average.pass.length.m + Accurate.long.pass.length.m*, *Passes.to.final.third.per.90 + Progressive.passes.per.90*, and *Accurate.long.pass.length.m + Progressive.passes.per.90* exhibit high absolute correlations, with values around 0.7 or higher. These pairs are likely to capture overlapping aspects of *passing OBV*, such as the effectiveness and accuracy of long passes and the ability to progress the ball towards the opponent's goal. For this reason, a select few variables, namely *Passes.per.90*, *Accurate.passes.percent*, *Forward.passes.per.90*, *Long.passes.per.90*, *Accurate.forward.passes.percent*, *Assists.per.90*, and *Key.passes.per.90*, have been removed from the dataset. *Passes.per.90* was removed because it is simply the sum of *Short.medium.passes.per.90* and *Long.passes.per.90*. *Forward.passes.per.90*, although likely a good predictor of *OBV*, shows strong correlations with and likely explains the same parts of *OBV* as *Progressive.passes.per.90* and *Passes.to.final.third.per.90*, which is why it was removed. *Long.passes.per.90* is another variable to be removed, as its influence is captured by *Average.pass.length.m* (0.826) and *Progressive.passes.per.90* (0.798). *Accurate.forward.passes.percent* is a variable highly correlated with *Progressive.passes.per.90* and *Passes.to.final.third.per.90*, which measure similar forward-passing actions. *Assists.per.90* is highly correlated with *Key.passes.per.90* and *xA.per.90*, both of which capture the contribution to goal-scoring opportunities, and finally *Key.passes.per.90* is removed as this is highly correlated (0.813) with *xA.per.90*.

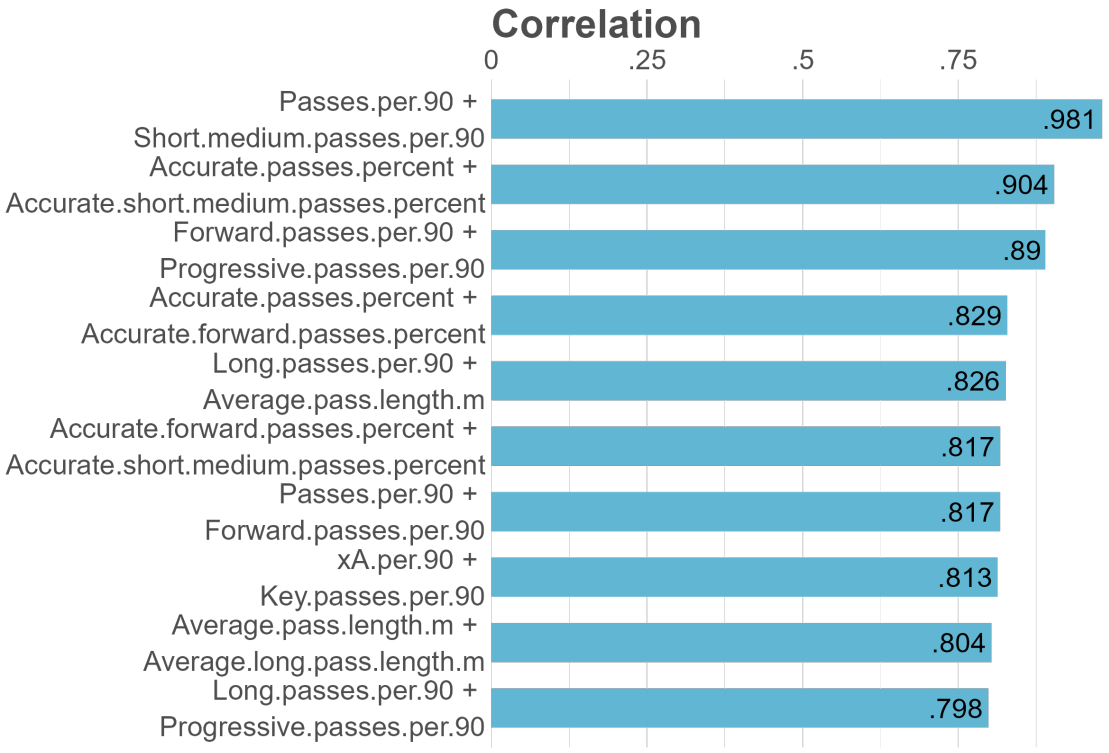

Figure A2: Passing variable correlations in Wyscout dataset

Figure A3 below shows the same observations except while focussing on the Wyscout defensive variables. Variables such as *Sliding.tackles.per.90 + PAdj.Sliding.tackles*, *Interceptions.per.90 + PAdj.Interceptions*, and *Successful.defensive.actions.per.90 + Interceptions.per.90* exhibit high absolute correlations, with values exceeding 0.9. These pairs likely capture overlapping aspects of defensive actions, indicating redundancy.

Based on this analysis, several variables were removed from the dataset due to high correlation with other input variables. *Sliding.tackles.per.90* shows an extremely high correlation (0.982) with *PAdj.Sliding.tackles*. Both metrics measure similar defensive actions (sliding tackles) but adjusted for different contexts. Removing *Sliding.tackles.per.90* helped avoid redundancy and allowed greater focus on the adjusted metric.

Interceptions.per.90 is highly correlated (0.955) with *PAdj.Interceptions*, which adjusts for possession. Since both metrics essentially measure interceptions but one

adjusts for possession to provide a more accurate reflection of a player's defensive performance, we retained *PAdj.Interceptions* and removed *Interceptions.per.90*.

Additionally, *Successful.defensive.actions.per.90* was removed due to its high correlation with multiple metrics, such as *Interceptions.per.90* and *Defensive.duels.per.90*. It overlapped significantly with other metrics that also reflect defensive success, and thus removing it helped maintain a clean dataset with less redundancy. *Successful.defensive.actions.per.90* was thus removed from the dataset.

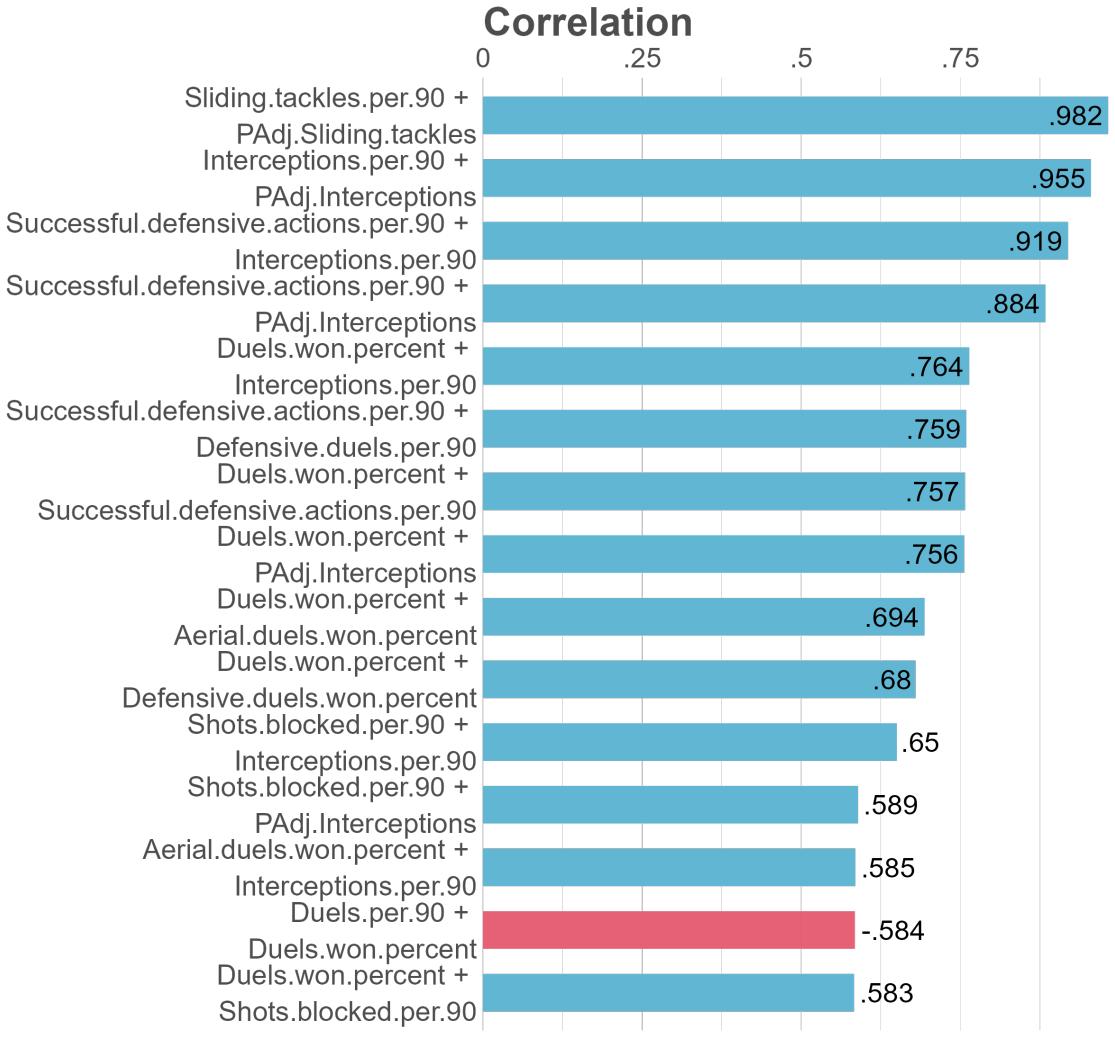

Figure A3: Defensive related variables correlation

Figure A4 below shows the correlation amongst dribbling related variables. It is quick to see that there are far less dribbling related variables than passing and defensive action related ones, suggesting that it may be difficult to model the dribbling component of *OBV*. Even though the number of accelerations and progressive runs are correlated, reducing the number of variables further may significantly hamper the model, and so all dribbling variables will be considered when modeling.

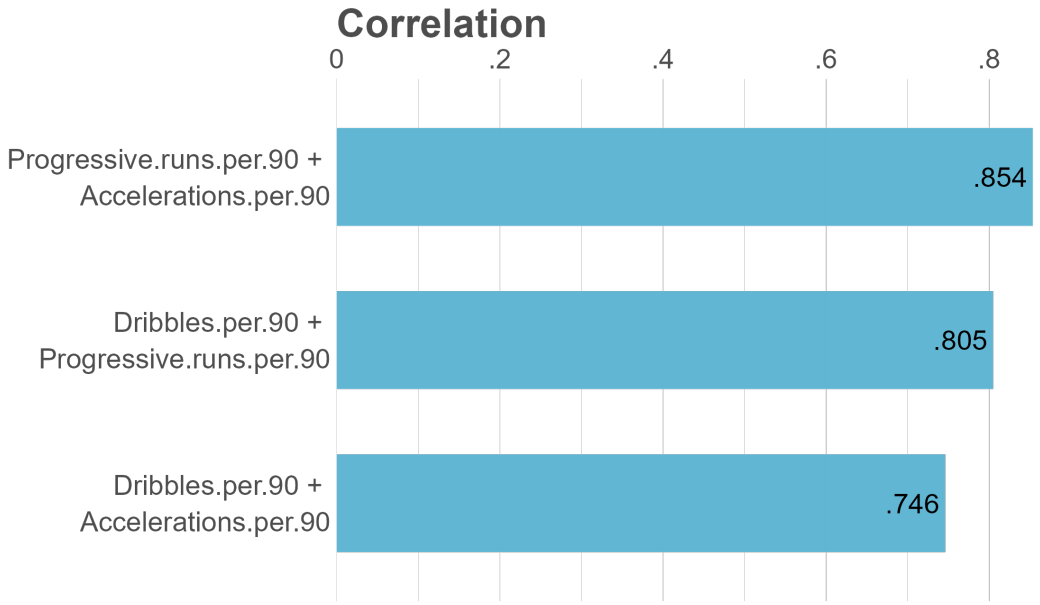

Figure A4: Dribbling related variables correlation

A.4.2 FBref Dataset

The *FBref* explanatory variables used in this analysis are shown in Tables A4,A5,and A6 in the appendix. These variables can once again be grouped into categories of passing, dribbling, and defending, to align with the three *OBV* components that will be modelled.

Passing metrics are extensive, detailing the player's distribution accuracy and effectiveness. Key variables include *Long.passes.per.90*, *Short.passes.per.90*, and *Ac-*

curate.short.passes.percent, which measure the frequency and accuracy of different types of passes. Specific attacking passing scenarios are captured with variables like *Progressive.passing.distance.per.90*, *Passes.into.final.3rd.per.90*, and *Passes.into.penalty.area.per.90*, reflecting the player's ability to advance the ball into critical attacking zones, and likely return a positive *Pass OBV* value. Additionally, metrics such as *Crosses.into.penalty.area* highlight the player's capability in creating very high *OBV* chances, as was seen in Figure 5.

Advanced metrics like *Accurate.long.passes.percent* and *Accurate.medium.passes.percent* indicate a player's proficiency in executing passes over varying distances. *Progressive.passes.till.final.3rd.per.90* focuses on their role in moving the ball towards the attacking third. In contrast, *Non.progressive.passing.distance.per.90* measures passing activities that do not advance the play, and likely return negative *OBV* values, ensuring a comprehensive evaluation of the player's passing ability.

Dribbling involves the player's ability to manoeuvre the ball and bypass opponents. Important variables include *Carries.per.90* and *Take.ons.attempted.per.90*, which quantify the frequency of carries and number of times a player attempts to dribble past an opposition player. The success rate of these actions is captured by *Take.ons.succeeded.per.90*, highlighting the player's effectiveness in dribbling past defenders.

Progressive.carries.distance.per.90 and *Progressive.carries.till.final.3rd.per.90* measure the player's capability to carry the ball forward and make advancements up the field. Metrics such as *Carries.into.final.3rd.per.90* and *Carries.into.penalty.area.per.90* are likely to be critical variables when modeling *DC OBV* as they focus on a player's success in penetrating critical attacking zones through dribbling and are expected to be highly positively correlated with *DC OBV*. Additionally, *Non.progressive.carries.distance.per.90* provides insight into carrying activities that do not significantly advance the play, allowing for a balanced analysis of a player's dribbling ability to be modelled.

The Defending variables evaluate a player's defensive prowess. Variables such as *Tackles.in.attacking.3rd.per.90*, *Tackles.in.defensive.3rd.per.90*, and *Tackles.in.midfield.3rd.per.90* assess the player's engagement and effectiveness in tackling across each major zone on the pitch. The variable *Interceptions.per.90* focuses on the player's ability to intercept passes, a crucial defensive skill. Additionally, metrics like *Yellow.cards.per.90* and *Red.cards.per.90* provide insight into the player's disciplinary record, reflecting their tendency to commit fouls and receive cards. While these variables are crucial for assessing a player's overall defensive behaviour and discipline, they may lack the locational granularity required to accurately evaluate *DA OBV*, as illustrated in Figure 7.

Once again, we first considered the passing-related correlations. Two key feature engineering operations were conducted. First, *Progressive.passing.distance.per.90* from *Total.passing.distance.per.90* was subtracted to create *Non.progressive.passing.distance.per.90*, which allowed us to drop *Total.passing.distance.per.90* from the dataset. Additionally, *Passes.into.final.third.per.90* was subtracted from *Progressive.passes.per.90* to create *Progressive.passes.until.final.third*. Figure A5 below shows the correlation amongst passing variables in the *FBref* dataset.

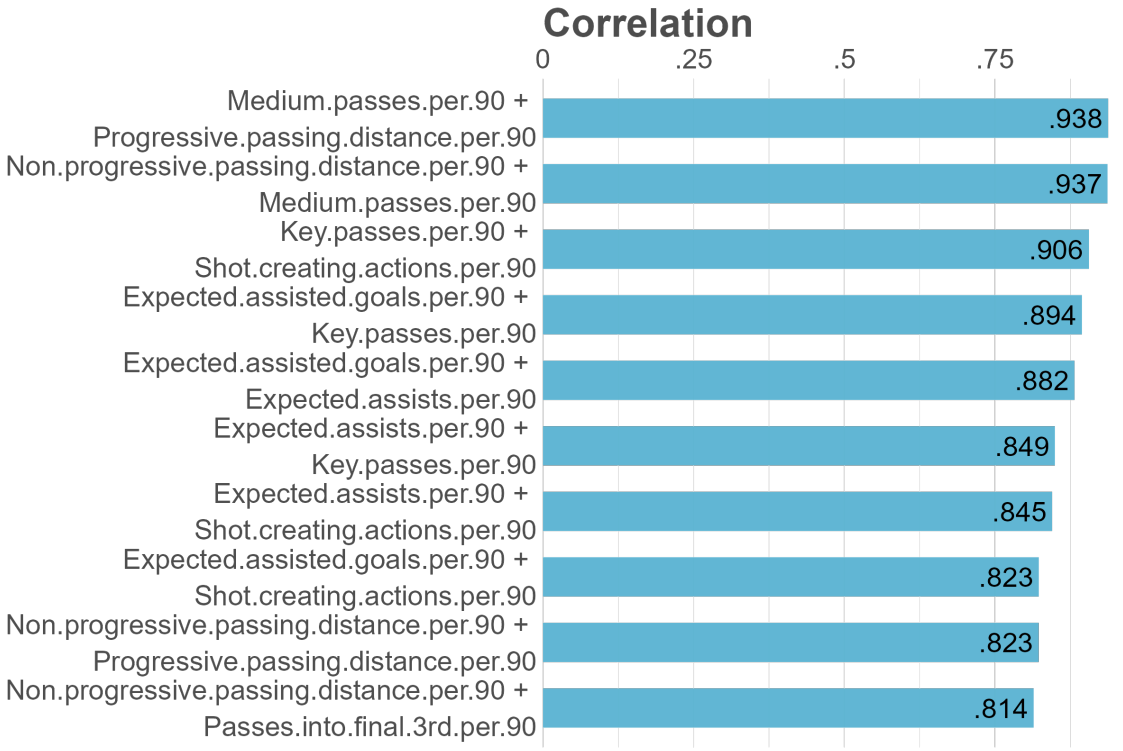

Figure A5: Correlation amongst passing variables in FBref dataset

Again, there were numerous passing variables which overlap in terms of high inter-correlations, and likely explainability of *Pass OBV*. This means variables needed to be removed. These variables included, *Medium.passes.per.90*, which was found to be highly correlated with *Progressive.passing.distance.per.90*, *Non.progressive.passing.distance.per.90*, and *Passes.into.final.third.per.90*, with the majority of the *OBV* explained by medium passes likely being captured by passes into the final third. *Shot.creating.actions.per.90* is also found to be highly correlated with several vari-

ables, such as *Passes.into.penalty.area.per.90*, *Key.passes.per.90* and *Expected.assists.per.90*. Naturally, it appears *Shot.creating.actions.per.90* is explained by the other variables well, and so it is also removed. *Expected.assisted.goals.per.90* is highly correlated with several variables, particularly *Expected.assists.per.90*, however, the latter is expected to produce a greater understanding of *OBV*, as it considers all passes within proximity of the box, whereas *Expected.assisted.goals.per.90* is specific to passes which directly lead to shots. *Key.passes.per.90* are simply passes which lead to shots, and thus are highly correlated with several variables, particularly *Expected.assists.per.90*, and is thus also removed. Finally, *Goal.creating.actions.per.90* has an expectedly high correlation with *Assists.per.90*, as there is understandably a large amount of correlation between assists and goal-creating actions, so goal-creating actions are removed due to unnecessary redundancy. Figure **??** highlights several defensive variables that have a weak to moderate correlation with each other.

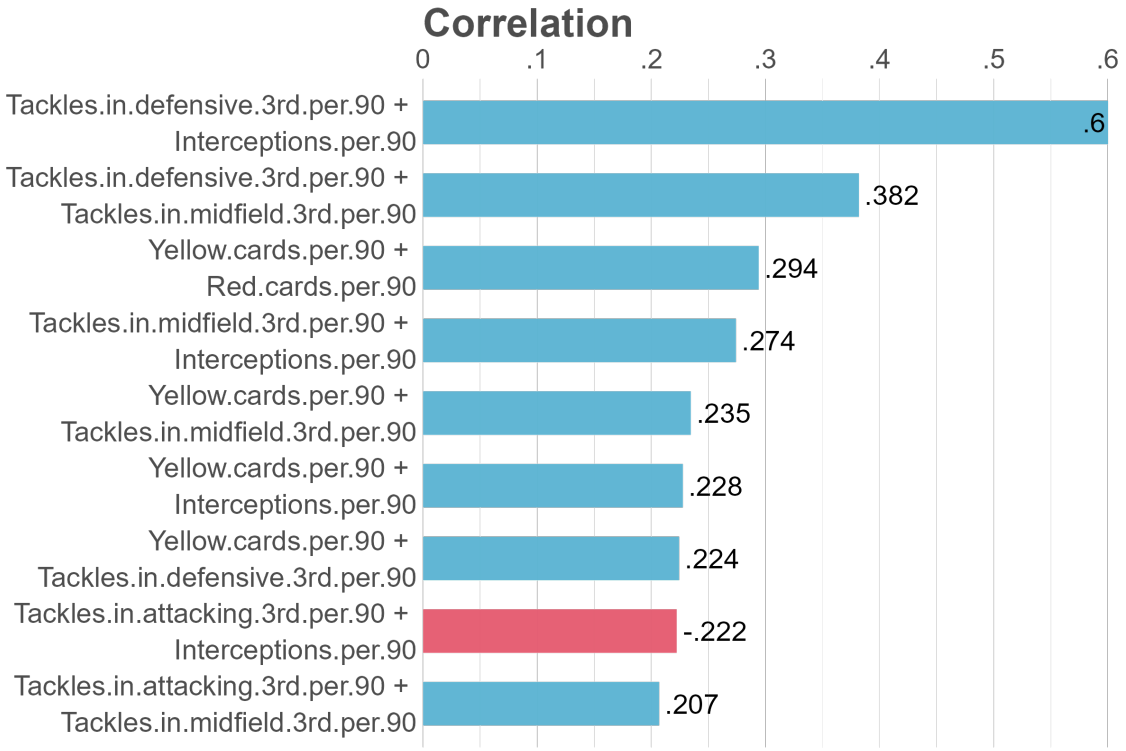

Figure A6: Correlation between defensive variables in FBref dataset

Tackles.in.the.defensive.3rd.per.90 combined with *Interceptions.per.90* show the high-

est correlation, with a correlation coefficient of 0.6. Since there were no strong intercorrelations present in the dataset, it was not necessary to remove variables due to multicollinearity.

The same feature engineering that was applied to the passing variables was also performed on the dribbling ones. Specifically, *Progressive.carries.distance.per.90* was subtracted from *Total.carries.distance.per.90* to create *Non.progressive.carries.distance.per.90*, which allowed us to drop *Total.carries.distance.per.90* from the dataset. Additionally, *Carries.into.final.3rd* was subtracted from *Progressive.carries.per.90* to create *Progressive.carries.until.final.3rd*. Figure A7 below shows the correlations between dribbling variables, revealing several notable relationships.

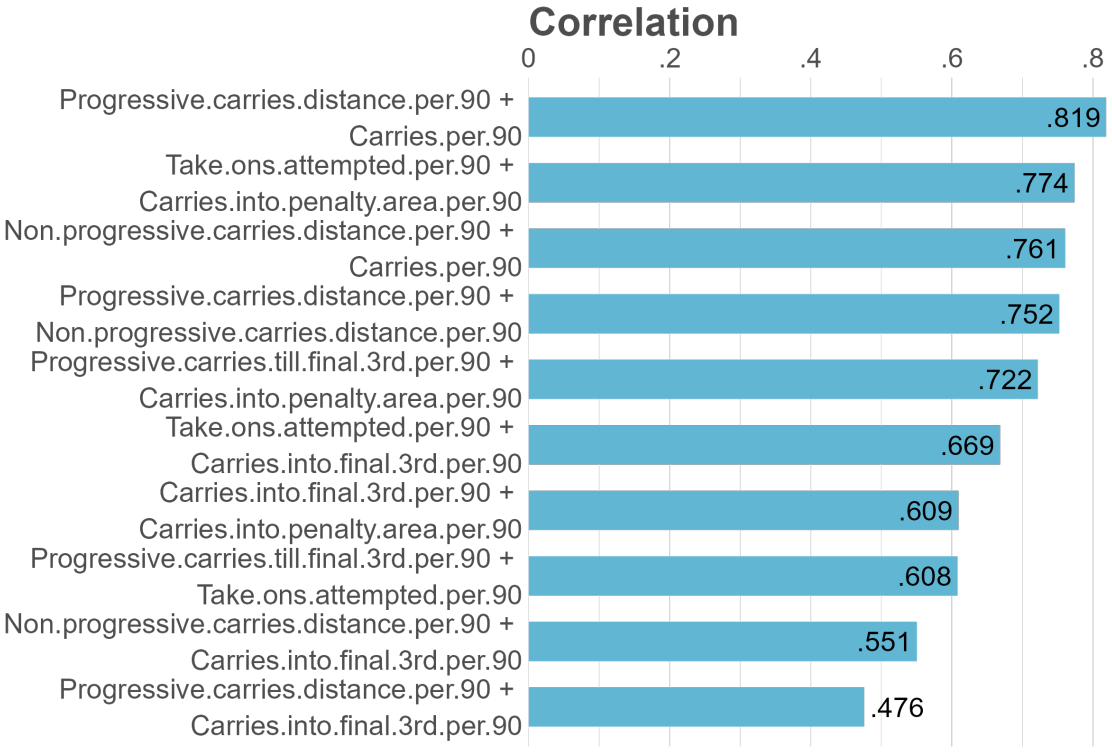

Figure A7: Correlation between dribbling variables in FBref dataset

The strongest correlation exists between *Progressive.carries.distance.per.90* and *Carries.per.90*, which is intuitive as players who carry the ball more frequently tend to cover greater progressive distances. However, these metrics capture distinct aspects of a player's ball progression abilities: *Carries.per.90* reflects how often a player

chooses to dribble, while *Progressive.carries.distance.per.90* indicates their effectiveness at advancing the ball upfield. Similarly, while *Take.ons.attempted.per.90* and *Carries.into.the.penalty.area.per.90* show considerable correlation, they measure different skills – the former indicates a player's willingness to challenge defenders directly, while the latter demonstrates their ability to penetrate high-value areas of the pitch.

Given that each variable provided unique insights into different facets of dribbling performance, all dribbling variables were retained in the analysis despite these correlations.

A.5 Principal Component Analysis (PCA)

A.5.1 Full Dataset

Variable	PC1	PC2
Defensive.duels.won.percent	-0.18	-0.08
Crosses.to.goalie.box.per.90	0.17	-0.19
Progressive.passes.per.90	-0.19	-0.30
Accurate.crosses.from.left.flank.percent	0.06	-0.06
Progressive.runs.per.90	0.21	-0.18
Fouls.per.90	0.07	0.09
Accurate.crosses.from.right.flank.percent	0.03	-0.08
Accurate.back.passes.percent	-0.10	-0.18
Aerial.duels.per.90	-0.02	0.31
Smart.passes.per.90	0.19	-0.15
Passes.to.penalty.area.per.90	0.18	-0.28
Accelerations.per.90	0.22	-0.15
Back.passes.per.90	0.20	-0.23
Accurate.long.pass.length.m	-0.23	-0.19
Crosses.from.left.flank.per.90	0.11	-0.14
Accurate.passes.to.penalty.area.percent	0.11	-0.07
Duels.won.percent	-0.28	-0.12
Crosses.from.right.flank.per.90	0.13	-0.14
Shots.blocked.per.90	-0.24	0.07
Average.pass.length.m	-0.23	-0.08
Duels.per.90	0.22	0.14
Accurate.crosses.percent	0.00	-0.09
Accurate.smart.passes.percent	0.06	-0.10
Dribbles.per.90	0.27	-0.07
Aerial.duels.won.percent	-0.23	0.01
Passes.to.final.third.per.90	-0.15	-0.29
xA.per.90	0.22	-0.16
Accurate.progressive.passes.percent	0.10	-0.22
Short.medium.passes.per.90	-0.12	-0.30
Defensive.duels.per.90	-0.09	-0.15
PAdj.Interceptions	-0.26	-0.11
Accurate.short.medium.passes.percent	-0.20	-0.20
Successful.dribbles.percent	-0.06	-0.08
PAdj.Sliding.tackles	-0.07	-0.06

Table A7: Wyscout Variable Loadings on the First Two Principal Components

Variable	**PC1**	**PC2**
Tackles.in.attacking.3rd.per.90	-0.13	0.10
Tackles.in.defensive.3rd.per.90	0.14	-0.02
Accurate.short.passes.percent	0.23	0.18
Passes.into.final.3rd.per.90	0.25	0.19
Short.passes.per.90	0.18	0.27
Progressive.passes.received.per.90	-0.28	0.19
Progressive.carries.till.final.3rd.per.90	-0.20	0.15
Progressive.passes.till.final.3rd.per.90	-0.21	0.17
Non.progressive.passing.distance.per.90	0.28	0.21
Tackles.in.midfield.3rd.per.90	0.05	0.04
Take.ons.succeeded.per.90	0.02	-0.05
Take.ons.attempted.per.90	-0.25	0.17
Carries.into.final.3rd.per.90	-0.14	0.28
Carries.per.90	0.22	0.29
Expected.assists.per.90	-0.16	0.27
Long.passes.Per.90	0.23	0.12
Accurate.medium.passes.percent	0.26	0.09
Crosses.into.penalty.area	-0.08	0.17
Assists.per.90	-0.17	0.17
Progressive.passing.distance.per.90	0.30	0.11
Assists.Minus.Expected.Assisted.goals.per.90	-0.03	-0.03
Red.cards.per.90	0.02	-0.03
Interceptions.per.90	0.22	-0.04
Carries.into.penalty.area.per.90	-0.25	0.16
Through.balls.per.90	-0.08	0.18
Non.progressive.carries.distance.per.90	0.07	0.32
Yellow.cards.per.90	0.06	-0.08
Accurate.long.passes.percent	0.08	0.09
Progressive.carries.distance.per.90	0.12	0.30
Passes.into.penalty.area.per.90	-0.14	0.29

Table A8: FBref Variable Loadings on the First Two Principal Components

A.5.2 Passing Dataset

Variable	PC1	PC2
Accurate.crosses.percent	0.043	-0.111
Crosses.from.left.flank.per.90	0.188	-0.089
Accurate.crosses.from.left.flank.percent	0.123	-0.056
Crosses.from.right.flank.per.90	0.220	-0.097
Accurate.crosses.from.right.flank.percent	0.080	-0.083
Crosses.to.goalie.box.per.90	0.293	-0.123
Back.passes.per.90	0.299	-0.210
Accurate.back.passes.percent	-0.117	-0.233
Short.medium.passes.per.90	-0.125	-0.390
Accurate.short.medium.passes.percent	-0.259	-0.273
Average.pass.length.m	-0.299	-0.137
Accurate.long.pass.length.m	-0.275	-0.270
xA.per.90	0.358	-0.105
Smart.passes.per.90	0.288	-0.142
Accurate.smart.passes.percent	0.113	-0.113
Passes.to.final.third.per.90	-0.156	-0.398
Accurate.passes.to.final.third.percent	0.065	-0.224
Passes.to.penalty.area.per.90	0.325	-0.250
Accurate.passes.to.penalty.area.percent	0.183	-0.051
Progressive.passes.per.90	-0.190	-0.377
Accurate.progressive.passes.percent	0.168	-0.263

Table A9: Wyscout Passing Variables on PC1 and PC2

Variable	**PC1**	**PC2**
Assists.per.90	-0.143	0.317
Assists.Minus.Expected.Assisted.goals.per.90	-0.055	-0.034
Expected.assists.per.90	-0.079	0.468
Long.passes.Per.90	0.304	0.112
Accurate.long.passes.percent	0.152	-0.008
Accurate.medium.passes.percent	0.354	-0.060
Short.passes.per.90	0.323	0.220
Accurate.short.passes.percent	0.351	0.071
Progressive.passing.distance.per.90	0.389	0.010
Passes.into.final.3rd.per.90	0.369	0.134
Passes.into.penalty.area.per.90	-0.055	0.489
Crosses.into.penalty.area	-0.039	0.330
Through.balls.per.90	-0.026	0.343
Non.progressive.passing.distance.per.90	0.418	0.078
Progressive.passes.till.final.3rd.per.90	-0.187	0.347

Table A10: FBref Passing Variables on PC1 and PC2

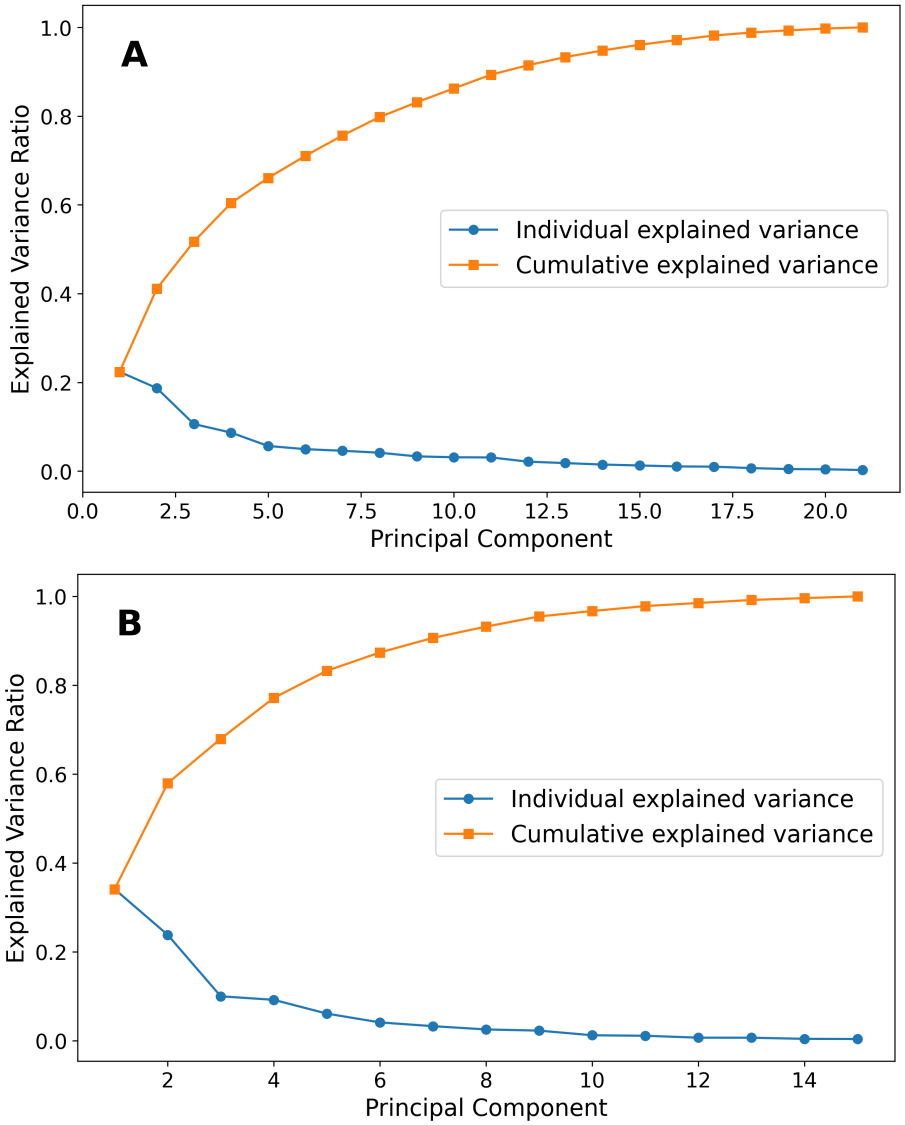

Figure A8: Screeplots of Wyscout (A) and FBref (B) passing datasets

A.5.3 Dribbling Dataset

Variable	PC1	PC2
Dribbles.per.90	0.565	0.029
Successful.dribbles.percent	-0.026	-0.998
Progressive.runs.per.90	0.590	-0.050
Accelerations.per.90	0.577	-0.022

Table A11: Loadings of Dribbling Variables on PC1 and PC2

Variable	PC1	PC2
Progressive.carries.till.final.3rd.per.90	0.336	-0.333
Progressive.carries.distance.per.90	0.379	0.407
Non.progressive.carries.distance.per.90	0.417	0.338
Take.ons.attempted.per.90	0.395	-0.363
Take.ons.succeeded.per.90	-0.038	0.018
Carries.into.final.3rd.per.90	0.441	-0.065
Carries.into.penalty.area.per.90	0.398	-0.379
Carries.per.90	0.243	0.574

Table A12: Loadings of FBref Dribbling variables on PC1 and PC2 (Rounded to Three Decimal Places)

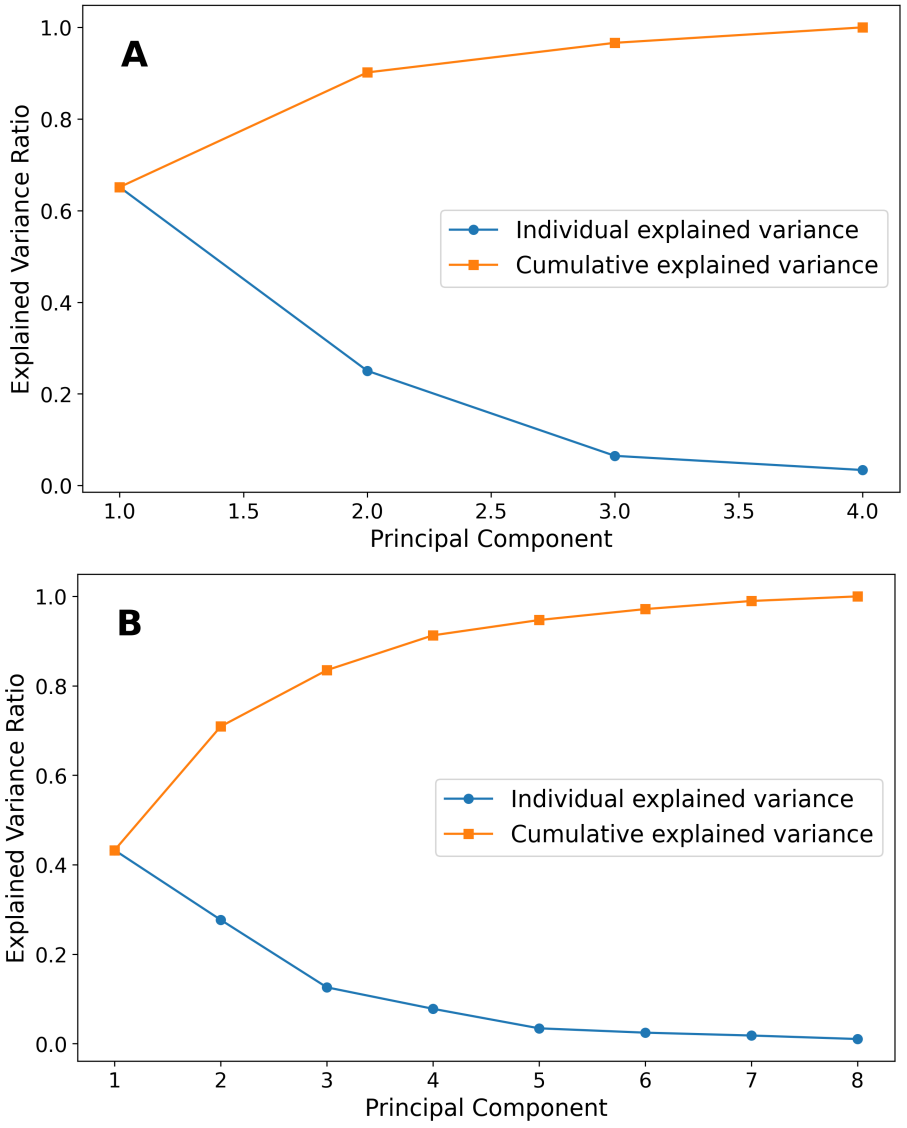

Figure A9: and FBref (B) dribbling datasets

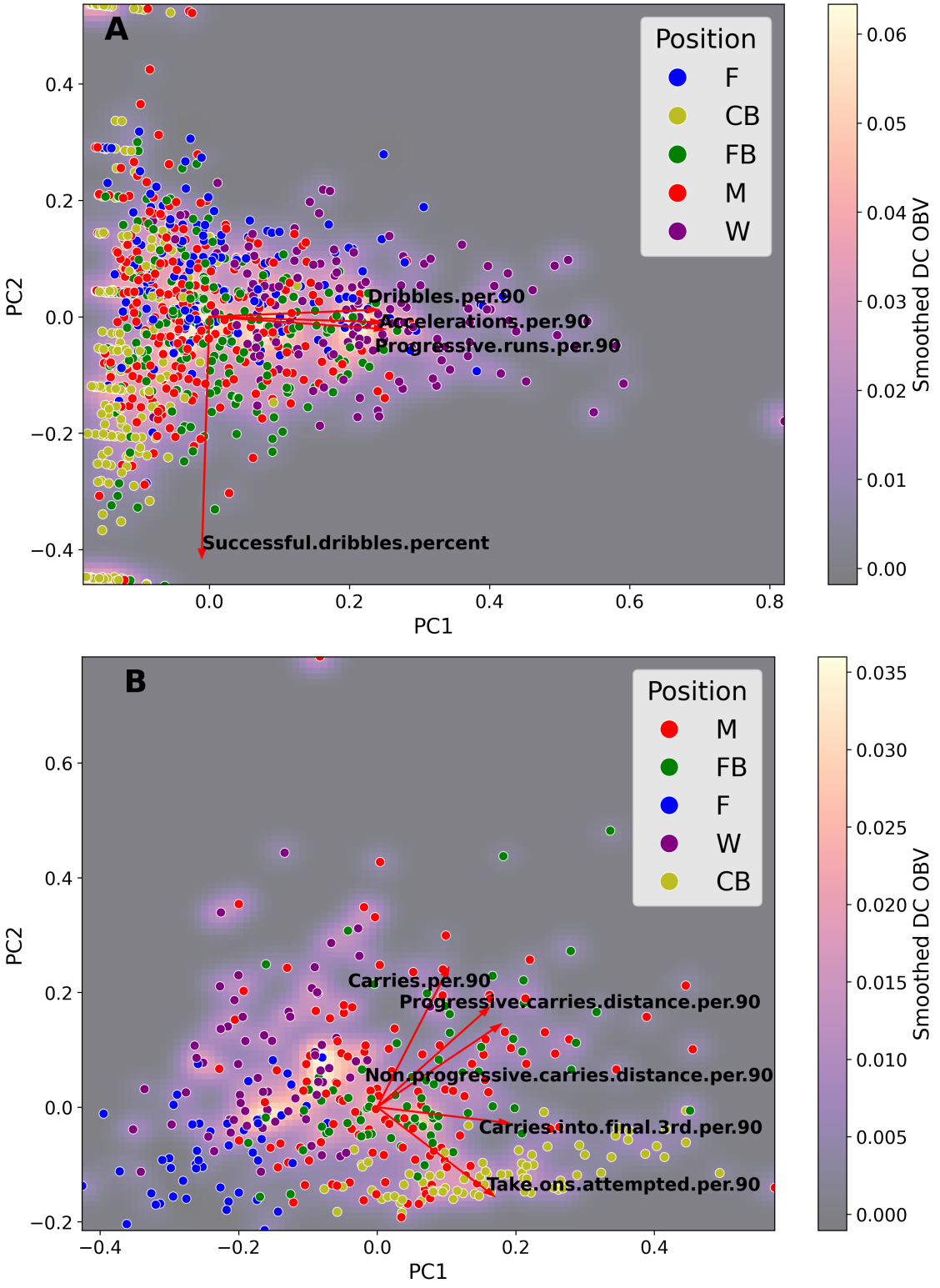

Figure A10: Biplots of Wyscout (A) and FBref (B) dribbling datasets

A.5.4 Defensive Action Dataset

Variable	PC1	PC2
Duels.per.90	0.346	0.223
Duels.won.percent	-0.473	0.016
Defensive.duels.per.90	-0.186	0.416
Defensive.duels.won.percent	-0.347	-0.170
Aerial.duels.per.90	0.078	0.069
Aerial.duels.won.percent	-0.356	0.088
PAdj.Sliding.tackles	-0.117	0.400
Shots.blocked.per.90	-0.359	-0.009
PAdj.Interceptions	-0.428	0.155
Fouls.per.90	0.202	0.522
Yellow.cards.per.90	-0.034	0.493
Red.cards.per.90	-0.022	0.189

Table A13: Wyscout Loadings of Defensive Variables on PC1 and PC2

Variable	PC1	PC2
Yellow.cards.per.90	0.389	0.352
Red.cards.per.90	0.192	0.268
Tackles.in.attacking.3rd.per.90	-0.041	0.743
Tackles.in.defensive.3rd.per.90	0.571	-0.174
Tackles.in.midfield.3rd.per.90	0.431	0.312
Interceptions.per.90	0.547	-0.352

Table A14: Loadings of Defensive Variables on PC1 and PC2 (Rounded to Three Decimal Places)

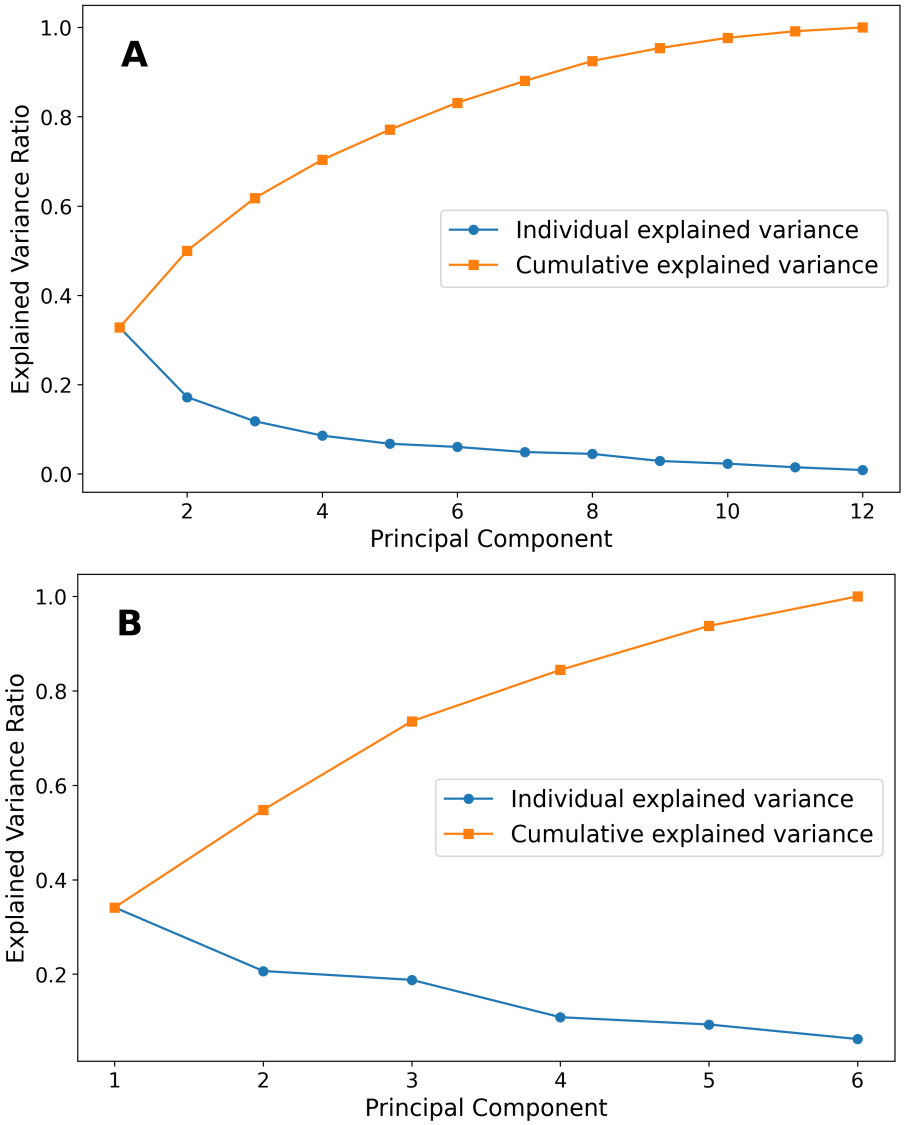

Figure A11: and FBref (B) defensive action datasets

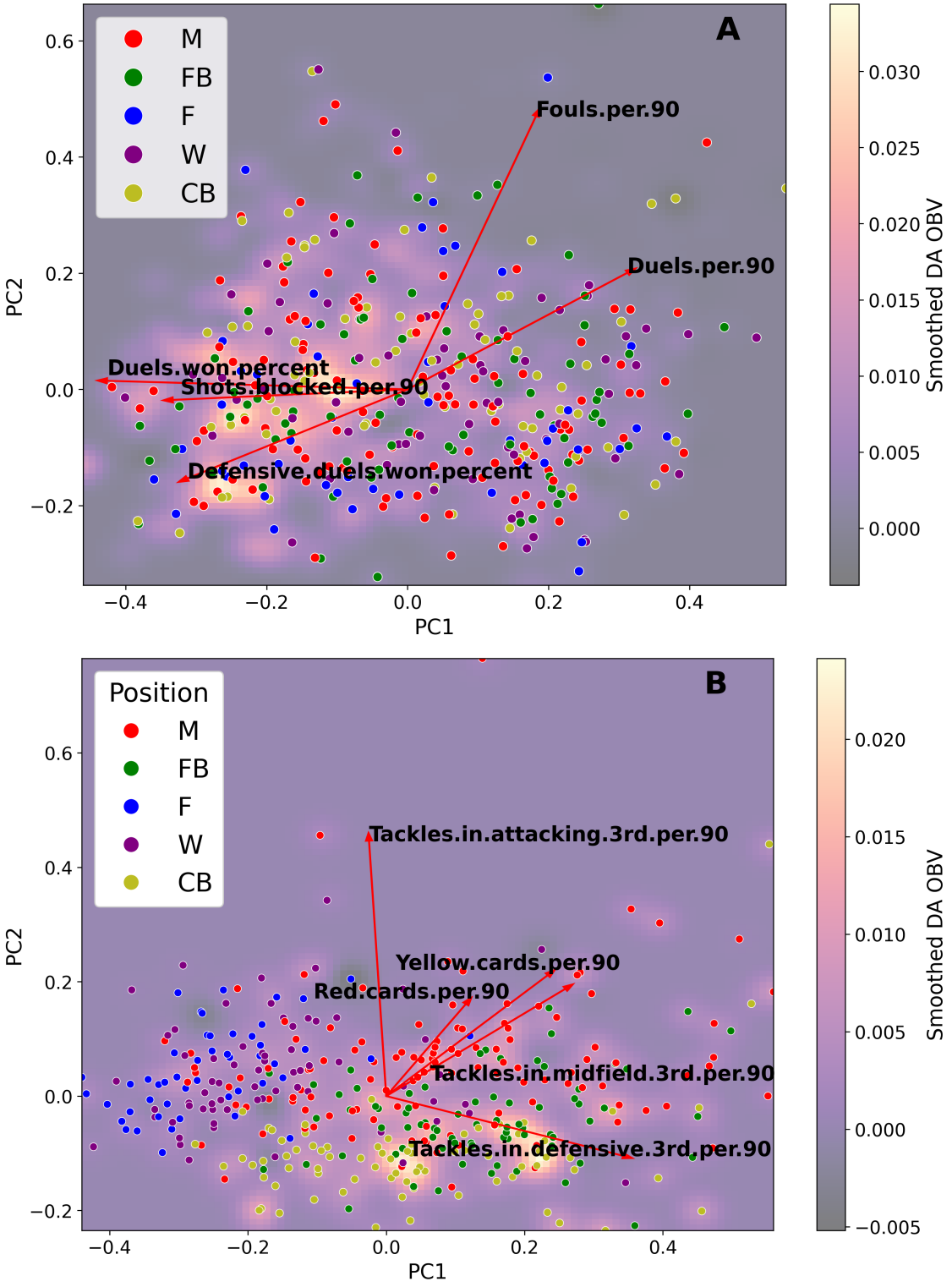

Figure A12: Biplots of Wyscout (A) and FBref (B) defensive action datasets