Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Business Analytics and Quantitative Marketing

Decision-Making Under Pressure: Quantifying Football Players' Passing Behaviour

Author:
Rohit Khiara
Student Number:
523302rk

Supervisor:
Michel van de Velden
Second Assessor:
NW Koning

Abstract

Prospect theory states that losses and gains are valued differently; this research aims to determine whether there is evidence for prospect theory in how football players and teams make passing decisions under pressure. In determining whether teams choose to minimise risk when making passing decisions, there are practical applications for teams whilst devising tactics. A novel dataset is used, which combines tracking and event data from the 2022 FIFA World Cup. The dataset records information for all actions on a football pitch, including the locations of the ball and players. Machine learning models are used to estimate the risk of a pass that is made under pressure, in addition to the risk of all other passes that could have been played in that situation - these are referred to as counterfactual passes. By dividing the pitch into zones, and based on the probability of a scoring action resulting from said zone, a value is assigned to the (counterfactual) passes. Two decision parameters, a *risk Decision Parameter* (*rDP*), and a *Risk-Value Efficiency* (*RVE*) parameter are introduced; for a given pass, these parameters evaluate whether a riskier option, or a more valuable option exist at the moment that a pass is made. The results are as follows: for a majority of the tournament, teams appear to prioritise minimising risk over finding valuable passes; teams that progress further in the tournament do not exhibit significantly different attitudes towards risk than other teams. Furthermore, teams demonstrate conservative behaviour consistent with prospect theory until the final stages of a tournament where it is deemed valuable to assume greater amounts of risk to achieve success. In addition, for games decided by penalty shootouts, teams that demonstrate more risk-seeking behaviour than their opponent during said games tend to achieve greater success in penalty shootouts.

February 2024

ERASMUS UNIVERSITEIT ROTTERDAM

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

I would like to thank my supervisor, dr. Michel van de Velden for his guidance and continued support throughout my thesis. A huge thank you to Angad and Yassine for accompanying me on this journey. Without you, none of this would have been possible or fun. Thank you to Shayna for always supporting and believing in me, and for watching football games with me that I claimed were for the purpose of “thesis research.” Last but not least, thank you to my parents for their sacrifice and unconditional love that has inspired me throughout the process.

Contents

1 Introduction

Traditional economic theory, namely, Expected Utility Theory, assumes that economic agents make decisions in a manner which maximises the utility or benefit that they gain from performing an action. Nevertheless, the field of behavioural economics has become increasingly prominent in recent decades, and has subsequently placed a greater emphasis on the fact that economic agents do not always act rationally or in their best interests. This notion is pioneered by prospect theory, introduced in Kahneman and Tversky (2013).

Fundamentally, prospect theory suggests that individuals value losses and gains differently. More specifically, prospect theory defines three biases that individuals are susceptible to in their decision-making: the certainty effect, the isolation effect, and loss-aversion. The certainty effect states that individuals tend to demonstrate risk-averse behaviour when faced with sure gains, and risk-seeking behaviour for choices that involve sure losses. Loss aversion hypothesises that individuals seek to minimise losses as opposed to maximising gains, even when the associated risk of a loss is minimal. Ultimately, prospect theory represents a more realistic modeling of the decision-making process than the expected utility theory.

This thesis will investigate the existence of the certainty effect and loss aversion in the decisions made by football players through their passing actions when under pressure. By researching prospect theory in the context of professional sports, the generalisability of this theory can be determined - that is, whether football players, and by extension, their teams value losses more than gains when faced with an increased risk of losing possession of the ball. Whereas prospect theory is founded on the idea of the predictable and known outcomes of a lottery, such an analysis allows for a more holistic understanding of prospect theory and whether it is applicable in a non-economic setting and a team dynamic as opposed to an individualistic setting. Additionally, analysing prospect theory in the context of football creates situations where, rather than looking at the decision-making process that is associated with a lottery, the outcome of a decision is not immediately known and the agents lack crucial information when making their choices.

Moreover, this thesis attempts to discover whether the nature of prospect theory is dynamic. In particular, investigating how the passing decisions of teams evolve over a certain period of games provides a framework to analyse whether teams are consistent with the decisions that they make in relatively similar situations across numerous games. That is, does the nature of the game, or the consequence of a reward alter a player's mental pressure and change the propensity for risk?

Thus, the extent to which specific components of prospect theory, such as the choice effect and loss-aversion, are situation-specific or constantly occurring can be established.

Beyond helping extend the plethora of academic literature and the existing knowledge on prospect theory, the research which is conducted in this thesis has practical relevance for the football industry. By obtaining a detailed understanding of players' psychology and their tendencies under pressure, there is a greater opportunity to utilise such analysis in the development of training methods and approaches to tactics. To this end, the following research question is investigated:

To what extent do football teams demonstrate risk-aversion in their passing behaviour when under pressure?

Risk-aversion is operationalised as the tendency of a player to frequently complete passes that are in a low-risk area of the pitch, whilst pressure refers to situations wherein an opposition player is within a certain distance of the ball and the player that is attempting the pass. The manner in which these factors are computed is briefly described below. To answer the research question, the data are obtained from StatsBomb (2021); the dataset consists of 360° event data for each game of various international tournaments and domestic league seasons. 360° data records every single event or action taken by a player in a given game, in addition to other variables such as the location of other players, the outcome of an action, the length and type of pass, and more.

For the purpose of this research, the data pertaining to the 2022 FIFA World Cup (Fédération Internationale de Football Association) is used, which consists of $234,626$ observations of 183 event-variables for the 64 games of the tournament. To ensure that the information that is extracted from the available data is maximised, the participating teams are grouped by the number of games they played in the tournament, and the risk-taking behaviour of teams is then analysed by group that their teams belong to. This approach results in five groups: teams that are knocked out in the Group Stage; Round of 16; Quarter-Finals; Semi-Finals; and Finals, corresponding to each round of the tournament. An additional group is included to account for the third and fourth placed teams that play an additional game against each other to determine their standing in the tournament. Analysing teams by groups can possibly help inform whether there are any commonalities in terms of risk-seeking behaviour between teams that are eliminated at identical stages of the tournament.

To conduct the analysis, firstly, the probability of a pass being intercepted is determined. Calculating this probability helps quantify the risk that the player assumes when passing the ball. In order to do so, two models are trained and tested - a logistic regression model, and a Random Forest model. Secondly, the value of the pass is obtained using an Expected Possession Value matrix which details the gain to a team of moving the ball from one location on the pitch to another. Using the model of choice, the risk of all possible counterfactual passes for a given pass is predicted, and its value is determined. Counterfactual passes refer to all other passes that could have been made at the moment that the original pass is made. Decision parameters are introduced to help quantify the decision-making of players and by extension, their teams, for passes made under pressure.

The main findings are as follows: teams seem to prioritise risk-minimisation over generating value in ball movement when faced with making passes under pressure; moreover, there is no significant difference in the risk-seeking behaviour of teams that progress from one round to the next relative to teams that are eliminated in a specific round. However, teams that do progress furthest in the tournament demonstrate a propensity for risk-seeking in the final stages relative to previous stages, which is consistent with prospect theory. Furthermore, teams that are able to make valuable passes under pressure do not necessarily need to assume a greater amount of risk to do so; these teams achieve a higher average ranking than teams that assume higher risk and make inefficient passes.

Accordingly, there are practical - tactical and strategic - contributions for teams when devising strategies for dealing with passes under pressure; theoretical contributions include providing a framework to analyse prospect theory in a sporting context, and supplementing existing knowledge of the performance of machine learning methods on imbalanced datasets.

The remainder of this paper is structured as follows: Section 2 considers the existing literature and approaches to quantifying football players' decision-making, whilst Section 3 describes the dataset in greater detail. Section 4 provides an extensive description of the methods that are used to answer the research question, and Section 5 outlines the results of conducting this research. Section 6 offers a summary of the work which is presented in this paper, whilst providing suggestions for future research.

2 Related Work

There is a considerable amount of literature dedicated to investigating prospect theory in a sporting context. Mundstock, da Silva Maia, and Bicalho (2021) hypothesises that, in the last 15 minutes of a game, football players take riskier decisions when faced with a loss as the match outcome, and finds that this is the case for the national football leagues of Brazil, England, and Germany. The study draws the conclusion that this behaviour is consistent with prospect theory, as the desire to avoid losing outweighs the desire to win. Therefore, teams often prefer to wait until the final moments of a game before taking riskier decisions that can change an adverse scoreline. This study presents one example of the manner in which prospect theory can, to some extent, explain the decisions that football players make, and how the game-state is a major factor in determining the risks that players are willing to take.

Similarly, Riedl, Heuer, and Strauss (2015) also examines whether prospect theory can be attributed to the decision-making of players and the strategies they employ, albeit in a different manner. Riedl et al. (2015) investigates the impact of FIFA's decision to increase the number of points that a team receives for winning a game from 2 to 3 to determine if there is any reduction in the amount of draws between teams. The number of expected draws is found by modeling the probability of a goal through a Poisson distribution, in order to estimate the individual results of the games that are considered in the study. The analysis is conducted over a period of 20 seasons for 24 countries, and finds that while the number of draws did show a slight decrease, 18% more matches ended in a draw than expected under the new regulations. It is concluded that, despite the presence of a greater reward for winning, prospect theory and loss-aversion maintain a dominating effect wherein the number of draws does not converge to the statistical expectation, as teams prefer to avoid losses. The work presented in Riedl et al. (2015) provides overarching support for the results that are found in Mundstock et al. (2021): whereas the latter focuses on and finds evidence of loss-aversion in specific time-periods of a game, the former shows that loss-aversion exists on a larger scale, and is persistent across multiple seasons.

Certain works have researched the existence of prospect theory in other sports; Bendickson, Solomon, and Fang (2017) explores its applications in the American National Football League (NFL). Bendickson et al. (2017) suggests that the NFL can be viewed as analogous to an organisation, such that theories of organisational behaviour can be applied. Accordingly, Bendickson et al. (2017) tests the following hypothesis: organisations will engage in risk-averse behaviour when performing above a certain reference point, and vice-versa. In the context of the NFL, pass attempts and interceptions are studied, to compare the proportion of interceptions when teams are winning and losing. The results indicate that teams are more likely to incur interceptions in the latter situation, thereby supporting the hypothesis and showing evidence of loss-aversion, as teams tend to take less risks having reached a certain reference point (a winning position). Thus, this study yields some evidence of the fact that loss-aversion in a sporting context is not limited to certain sports, but is rather widespread and recurring.

It holds that the academic literature with respect to prospect theory in sports is relatively well-documented. Nevertheless, there is a dearth of research which considers the influence of prospect theory on players' decision-making in international football tournaments. Whilst the research above considers domestic league football, there are several facets of international football tournaments such as the World Cup which significantly differ from the aforementioned club football. For instance, whereas teams compete in a domestic league every calendar year, the World Cup occurs once every four calendar years. As a result, many players do not get the chance to compete in multiple World Cups in their careers; in fact, fewer than 100 players have participated in four or more World Cups (Wikipedia contributors, 2024). Accordingly, the stakes of participating in a given World Cup tournament may possibly exceed that of a domestic league season, given the latter's annually recurring nature. Consequently, players may be faced with a greater pressure to perform, given a chance to represent their countries at what is arguably the pinnacle of professional football.

Moreover, an additional aspect of international football which deviates from club football is in terms of training. Players do not have the opportunity to train with their international teams in a given year as frequently as they do with their club teams. Therefore, it is likely that players are less familiar with the style and mannerisms of their international teammates relative to their club teammates. Moore, Adams, O'Dwyer, Steel, and Cobley (2017) corroborates this idea by inspecting whether factors such as team familiarity are significant in successfully identifying the preferred kicking-foot of Australian football players. The outcomes of this research demonstrate that players are better able to recognise the kicking-foot of their teammates than opposition players, in terms of accuracy as well as speed. It holds that players are more aware of the play-style of those players that they train with more frequently, which is likely to influence the decisions they make on the pitch.

Furthermore, a plethora of research has investigated and found evidence in support of the existence of home advantages in football (Peeters & van Ours, 2021; Pollard, 2006), which substantiates the importance of familiarity in the decision-making process by football players, and by extension, the outcomes of football games.

Therefore, accounting for the differences between international football and club football is necessary when studying prospect theory within the decision-making by football players. Based on the literature that is assessed above, the following sub-question is considered:

How does the risk-seeking behaviour of teams change at different stages of the World Cup?

It can be argued that as the players play more games with their international team at the tournament and progress further, they become more familiar with their teammates and roles, which can help exacerbate their propensity for risk. However, there is an overwhelming amount of literature which suggests that loss-aversion is a highly dominant effect. Thus, the following hypothesis is formulated:

Teams will become increasingly conservative with their passes as they progress through the World Cup.

A secondary sub-question is developed to investigate the relationship between risk-seeking behaviour and success:

Do teams that demonstrate a greater propensity for risk-seeking passes under pressure achieve a better rank in the World Cup?

This sub-questions aims to detail whether teams and players that do not succumb to loss-aversion tend to achieve greater success, as they are likely to approach their matches with a contrasting, more aggressive mentality to other teams. The following hypothesis is proposed:

Teams that are more risk-seeking for passes under pressure will have a higher rank in the World Cup relative to teams that are less risk-seeking.

Football has been relatively slow in its adoption of analytics and a data-driven approach to the sport compared to other international sports leagues such as those in North America. Nonetheless, research performed by Cefis (2022), which reviews academic publications related to football over the last decade, shows that articles with key words such as *machine learning* and *artificial intelligence* have become increasingly widespread since 2016. Thus, whilst analytics is still a relatively novel field with respect to football literature, it is showing signs of growth.

To this end, several papers have attempted to investigate the validity of metrics such as *expected goals* (xG), which, according to Mead, O'Hare, and McMenemy (2023), has become almost synonymous with football analytics given the extent of its popularity. The prominence of xG as a metric is expected because football, at its foundation, is a goal-oriented sport - therefore, quantifying shots and understanding more about the quality of a shot is of particular interests to the stakeholders of the football industry. Nonetheless, Brechot and Flepp (2020) claims that despite the ubiquity of xG, it has not been thoroughly assessed through academic literature. The same follows for the quality of a pass in terms of its contribution towards scoring a goal, or the decisions that players make when passing the ball, in terms of value added and risk. Hence, this thesis will supplement the existing academic literature in relation to passes and the decision-making process by providing an amalgamation of a statistical and behavioural-economic framework as a lens through which these topics can be studied. Thus, a critical evaluation of the current literature is pertinent to this section.

Burriel and Buldú (2021) develops a "minimal model" using the StatsBomb 360° data to examine the decision-making of football players as they make a pass. The approach focuses on 37 matches played by FC Barcelona in the $2020-2021$ La Liga season - Burriel and Buldú (2021) serves as a starting point for the methods that are implemented in this thesis. Burriel and Buldú (2021) relies on the use of the *Expected Possession Value* (EPV) parameter to quantify the benefit of a given pass. However, the EPV is not obtained from the StatsBomb 360° data in the methods used in Burriel and Buldú (2021). Rather, it is extracted using a historical dataset that contains relevant information for multiple teams across numerous seasons. As a result, the matrix of EPV values is the same for all teams. That is, it assumes that regardless of playing styles or players, each team generates the same value or benefit from a pass on the pitch. Thus, the EPV values that are used in Burriel and Buldú (2021) are not specific to FC Barcelona, thereby reducing the validity of the analysis, whilst facilitating greater computational feasibility as it is not required to calculate the team-specific EPV.

However, while it is disadvantageous that the EPV is a generalised measure of value, certain aspects of this thesis negate a variety of these issues. Firstly, as the EPV matrix is computed based on ball events from various European leagues, it is likely to be appropriate to a certain extent for the teams which participate in the World Cup for the following reason: it is plausible that the zone values are likely to capture the diversity in playing styles and tactics across leagues, which is useful when considering the fact that the World Cup is an international tournament, where players across leagues represent their countries. Thus, to some extent, the EPV measure may provide realistic zone values, highlighting why it is the preferred option in this work. Secondly, the World Cup provides a sample of 64 games. This sample size is insufficient to use for the computation of a team-specific metric, as each team plays between three to seven games. The EPV on the other hand, as previously mentioned, is based on numerous seasons worth of data. Therefore, it is probable that the EPV will serve as a relatively more accurate representation of the value of a pass than any metric which can be computed using the StatsBomb 360° data.

To distinguish itself from the approach in Burriel and Buldú (2021), this thesis opts for a different method of creating a model to quantify passes. The main divergence occurs in modelling the risk of a pass. Whereas Burriel and Buldú (2021) uses a physics-based framework to obtain the probability of a pass being intercepted by the opposing team or its own team, this thesis relies on machine learning techniques to predict the probabilities. This difference in approach is motivated by the fact that the methodology in Burriel and Buldú (2021) employs tracking data which is specific to passes in La Liga; this thesis aims to assess the decision-making exclusively through StatsBomb's readily-available, open-source 360° data as it enriches event data with freeze-frames that record the positions of all the players that are visible within the broadcast camera of a given match. Therefore, an alternative means of measuring the risk of a pass is required, one which can be performed through solely using the StatsBomb 360° data. Accordingly, this thesis supplements existing literature by demonstrating the scope of analysis that is possible with this dataset.

The use of machine learning to help value the decision-making of football players is not a novel feature of this thesis. Pulis and Bajada (2022) creates a "Decision Value" metric that uses Reinforcement Learning to value player actions. This metric takes into account the positions of teammates and opponents by utilising a combination of event and tracking data. The role of counterfactual situations is briefly described in Section 1, wherein a machine learning model is responsible for predicting the risk of passes that *could have* occurred. Existing literature relating to the analysis of counterfactual situations, namely Van Roy, Robberechts, Yang, De Raedt, and Davis (2021), proposes a Markov Decision Process to model a team's behaviour along with probabilistic model checking tools that result from Artificial Intelligence. Research conducted by García-Aliaga, Marquina, Coteron, Rodriguez-Gonzalez, and Luengo-Sanchez (2021) demonstrates that it is possible to use a combination of player actions and machine learning to classify players into their positions or roles on the pitch; it holds that machine learning is capable of more than just valuing the actions of players - it can also use that information to assess the manner in which players play. It holds that machine learning is an appropriate tool for this research, given the variety of its applications in football analytics. It is crucial to note that the problem of classifying passes as intercepted or not is highly imbalanced because the former situation is much less frequently occurring in a game relative to the number of passes that are completed.

Existing research has additionally attempted to quantify the psychological phenomenon of pressure (and inherently, decision-making) in football using machine learning. Bransen, Robberechts, Van Haaren, and Davis (2019) creates a framework to measure the mental pressure experienced by players before and during a game, and its impact on their performance. This analysis is conducted with a tree-based method. In fact, tree-based machine learning methods are frequently utilised in the literature; for instance, Van Roy et al. (2021) implements Gradient Boosted Trees Ensembles to model football players' unsuccessful actions. Trees are also applied in other sporting contexts; to retrieve similar plays in basketball from multi-agent spatiotemporal tracking data, Sha et al. (2017) uses a novel tree-based alignment method which facilitates a comparison between the similarity of the data.

Moreover, tree-based methods such as Random Forests are also used in abundance for the purpose of predicting probabilities in sports. Lock and Nettleton (2014) estimates the Win Probability prior to any play of a given NFL game through a Random Forest method. Furthermore, Schauberger and Groll (2018) shows that Random Forests generally outperform regression models in predicting the outcomes and number of goals for football matches in the World Cup tournaments hosted between $2002 - 2014$. Ali, Khan, Ahmad, and Maqsood (2012) additionally argues that Random Forests are easily interpretable amongst popular machine learning methods.

Accordingly, this thesis implements a Random Forest to predict the probability of a pass being intercepted. As noted earlier, imbalanced data presents a challenge in this thesis. Ruiz and Villa (2008) evaluates two methods of supervised learning, logistic regressions and Random Forests, on imbalanced meteorological data. The results of the evaluation indicate that both methods provide comparable results, whilst the logistic regression maintains a faster performance and simpler interpretation of the explanatory variables. On the other hand, Muchlinski, Siroky, He, and Kocher (2016) applies both approaches on an imbalanced dataset pertaining to civil war onset and concludes that Random Forests outperform logistic regression models in terms of predictive performance. More specifically, Muchlinski et al. (2016) finds that the Random Forest beats a logistic regression, in addition to a logistic regression model that corrects for imbalanced data.

As a result, this thesis implements an additional model to predict interception probabilities: a logistic regression. The goal is to examine the performance of the logistic regression model and Random Forest model on the imbalanced and balanced data, and contribute to the existing field of research with regards to their capabilities and versatility. Another sub-question follows:

Does a logistic regression model offer better prediction performance than a Random Forest model on imbalanced data or balanced data?

Based on the literature that is explored above, a corresponding hypothesis is formulated:

A Random Forest model will outperform a logistic regression model on both, balanced and *imbalanced data in predicting interception probabilities.*

The manner in which this hypothesis is tested is described further in Section 4.

3 Data

The data for this research are obtained from StatsBomb (2021). This specific dataset is chosen due to certain characteristics, namely, the availability of 360° data. Consequently, every event in the dataset contains a freeze-frame which shows the location of all players within the range of the camera that broadcasts the game. Therefore, the 360° data serves as a combination of tracking and event data; StatsBomb has made such data available for three competitions: UEFA Euro 2020, UEFA Women's Euro 2022, and the FIFA International World Cup 2022 (WC). The latter competition is selected for further analysis in this research, and is divided into two distinct categories: matches, and events.

There are 32 teams in the WC, and 64 matches are played over the duration of the tournament. For each match a variety of information is provided such as the score, referee, stadium name, and stage of the competition. Note that in the WC there are four stages: the Group Stage, Round of 16, Quarter-Finals, Semi-Finals, and Finals. The two teams that do not proceed beyond the Semi-Finals play an additional game against each other to determine which one of them places third and fourth in the tournament, in what is known as the 3$^{\text{rd}}$ Place Final.

Events, on the other hand, refer to every single action that occurs on the pitch during a game. For the 64 games of the WC, there are a total of $234,626$ events recorded for 187 variables such as the time of the action, passes, player locations, pass outcome, pass end locations, and more. A complete list of all the variables can be found in the StatsBomb Open Data Specification (2019). This data specification notes that in the dataset, on average, 28.71% of the events are passes - the most of any event.

As the analysis in this research is focused exclusively on passing behaviour, only pass events are taken into consideration. This analysis focuses on passes made during open play - defined as periods of normal play where a passing event does not occur from a situation where play is resumed following the ball going out-of-bounds or foul event. As per the StatsBomb Open Data Specification (2019), a pass can be classified according to the following types: "NA," "Kick Off," "Corner," "Throw-in," "Goal Kick," "Interception," and "Free Kick", where "NA" denotes a pass during regular play. To ensure that only passes that occur during open-play are taken into consideration, passes with the types “Kick Off,” “Corner,” “Throw-in,” “Goal Kick,” and “Free Kick” are removed from the dataset. Ultimately, there are $53,990$ unique passes in the resulting dataset that is used in the analysis.

There are numerous variables that are associated with a pass; the ones which are selected in this research are displayed in Table 1.

Table 1: Selected pass variables and their definitions as given in StatsBomb Open Data Specification (2019).

Variable	Details
id	A unique identifier for each pass.
duration	The length in seconds of the pass.
$pass.length$	The length in yards of the pass, from its origin to its destination.
$pass.angle$	The angle of the pass, in radians.
$location.x$	The x-coordinate at which the pass originated.
$location.y$	The y-coordinate at which the pass originated.
$ff\_location.x$	The x-coordinate of a player that is visible during the pass.
$ff\_location.y$	The y-coordinate of a player that is visible during the pass
$end\_location.x$	The x-coordinate of where the pass ends.
$end\_location.y$	The y-coordinate of where the pass ends.

Other pass variables such as $pass.height$, which specifies whether a pass is a ground, low, or high pass, are informative in the context of this research. However, their inclusion relies on making assumptions that cannot be proven: for the counterfactual passes that will be considered in the analysis, it is uncertain with which height the player would pass the ball to another player - as it is a counterfactual scenario. Therefore, to avoid making assumptions with respect to the nature of the pass, only variables that can be determined with certainty are utilised.

In order to assess the counterfactual scenarios of each pass, an additional variable, *freeze_frame* is considered. For each pass, this variable contains the locations of all players that are visible in the broadcast camera during the event. Therefore, each of the $53,990$ passes contain a *freeze_frame*. Accordingly, the freeze frames are extracted for each pass as a row, such that each unique pass-id consists of multiple rows in the dataset, where each row corresponds to a player that is associated with that specific pass. It holds that for a unique pass-id, its number of rows are determined by the number of players that are seen in the broadcast camera at the moment that the pass is played by the actor. Simply stated, one row records the original pass, and the other rows record the position of each player that is visible during the event - therefore, a pass with nine broadcasted players will have more rows than a pass with five broadcasted players.

Additionally, *freeze_frame* records three boolean variables that indicate whether the player is an *actor* - the player performing the pass; *teammate* - a player belonging to the same team as the actor; or a *keeper* - a player whose role is the goalkeeper of either team. Passes for which *freeze_frame* is unavailable are excluded from the analysis on account of the fact that the (counterfactual) analysis cannot be conducted upon them, as the locations of the players for that specific pass are unknown.

From the variables that are listed in Table 1, excluding *id*, all of their values can be determined for a counterfactual pass given information such as Euclidean coordinates. The exception is for the *duration* variable that is associated with the pass. That is, it is unknown for a hypothetical pass how long it would take to reach its destination. Therefore, to perform the counterfactual analysis, it is necessary to predict this value for each counterfactual pass. Accordingly, this variable is studied in greater detail.

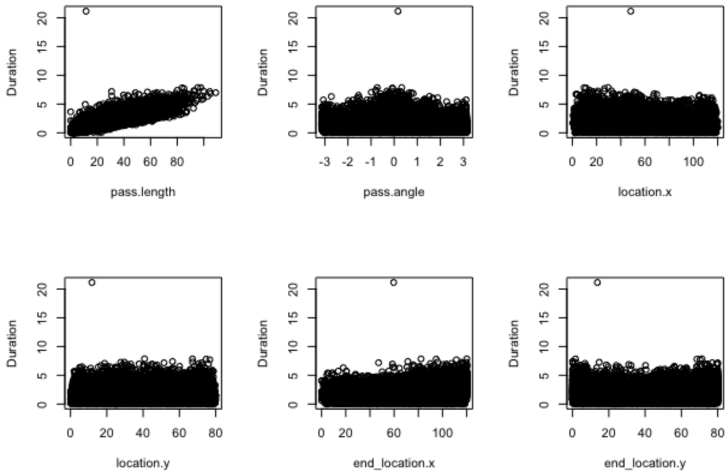


Figure 1: Scatter-plots of *duration* against its predictors

For each unique pass that is considered in the analysis, its value of *duration* is extracted. To obtain a more nuanced understanding of the relationship between *duration* and its predictors, consider the scatter-plots in Figure 1. Whilst the plot shows that there may be a linear relationship between the length of a pass and its duration, this relationship is not maintained for any of the other variables. It follows that in order to estimate the *duration* values of counterfactual passes, non-linear or non-parametric methods may need to be considered in Section 4.

Moreover, as previously discussed, risk-taking and pressure are central to this research. It is of utmost importance to focus on those passes that are played under pressure. Specifically, for each event, the variable *under_pressure* is recorded, which takes the value *TRUE* if the duration of a pressure event plus its timestamp overlaps with the duration of the pass, and takes the value *NA* otherwise. A pressure event is defined as one where a player closes down or blocks a passing lane. There are a total of $6,927$ passes from the $53,990$ which meet this criteria. Table 2 provides a breakdown of the number of passes made by each team, and the number of passes made under pressure.

Table 2: Number of total passes made and number of passes made under pressure for each team considered in the analysis

Team	**Number of Passes**		**Team**	**Number of Passes**	
	All	Under Pressure (%)		All	Under Pressure (%)
Argentina	3397	416 (12.25%)	Mexico	1044	113 (10.83%)
Australia	1278	218 (17.06%)	Morocco	2184	277 (12.68%)
Belgium	1469	200 (13.62%)	Netherlands	2424	330 (13.60%)
Brazil	2508	256 (10.21%)	Poland	1169	155 (13.26%)
Cameroon	968	184 (19.01%)	Portugal	2494	287 (11.52%)
Canada	1205	144 (11.95%)	Qatar	1132	177 (15.64%)
Costa Rica	835	151 (18.08%)	Saudi Arabia	893	140 (15.68%)
Croatia	3629	482 (13.29%)	Senegal	1283	196 (15.26%)
Denmark	1526	151 (9.89%)	Serbia	1185	156 (13.16%)
Ecuador	1158	157 (13.55%)	South Korea	1606	199 (12.40%)
England	2693	322 (11.97%)	Spain	3388	355 (10.48%)
France	2948	298 (10.11%)	Switzerland	1557	225 (14.45%)
Germany	1641	197 (12.00%)	Tunisia	1041	127 (12.20%)
Ghana	968	122 (12.61%)	United States	1792	180 (10.05%)
Iran	877	143 (16.30%)	Uruguay	1200	144 (12.00%)
Japan	1437	245 (17.03%)	Wales	1061	180 (16.97%)

4 Methodology

There are three main steps that are performed in the analysis: calculating the probability of the ball being intercepted, valuing the pass that is made, and lastly, quantifying the decision-making of the player.

To obtain the probability of a pass being intercepted, two approaches are compared - logistic regression models, and Random Forests. For each pass, these models predicts the probability of the surrounding opponents intercepting the ball before combining these values into a singular probability of the pass being intercepted.

The value of a pass, that is, its impact in terms of helping a team score a goal is measured through its *Zone Possession Value* (ZPV). ZPV is a metric which uses Friends of Tracking Data (2021) to determine the change in the *Expected Possession Value* (EPV) of the ball moving from one location of the pitch to another when a pass is made.

The above concepts are combined into two decision parameters - a *risk Decision Parameter* (*rDP*), and a *Risk-Value Efficiency Parameter* (*RVE*) - that quantify and evaluate the decision-making of a team. These decision parameters represent how much risk a team assumes with their passes over the tournament, and the value of these passes relative to the risk, respectively. The following subsections outline these processes in extensive detail.

4.1 Intercepting a Pass

Fundamentally, a pass occurs when a player sends the ball to the feet (or any body part which can legally be used to control the ball) of their teammate, or into some space which said teammate moves to occupy. An interception is said to have occurred when a player from an opposing team is able to successfully win control of the ball before it reaches its intended target or a player from the actor's team.

Analogously, Burriel and Buldú (2021) treats a player performing a successful pass to a teammate as an interception by that teammate, as they retain possession of the ball. These concepts of interceptions are used as a proxy measure of the risk that is assumed by the player when passing the ball. That is, a pass which has a high probability of being intercepted by an opponent instead of a teammate would entail a greater risk than vice-versa.

Predicting the probability of a pass being intercepted by its opponents will help in determining risk, one of the crucial components in quantifying the decision-making of players. To this end, the manner in which the probabilities of opponents intercepting the ball is elaborated upon below.

4.1.1 Defining Interceptions

To obtain the probability of a pass being intercepted by an opponent several factors are taken into consideration, the most important ones being the starting and ending locations of the pass and the surrounding opponents. Figure 2 displays an example of a passing situation wherein the actor attempts a pass; each of the opponents can move to intercept the ball - that is, each opponent player within the frame of the pass has some non-zero probability of regaining possession of the ball. Note that the recipient of the pass is not shown; the purpose of Figure 2 is to simply demonstrate the situation that is being considered - actors and opponents. For a given pass, it is therefore required to calculate the probability of each opponent intercepting the ball, and then combining these probabilities into a singular value that represents the probability of the pass being intercepted.

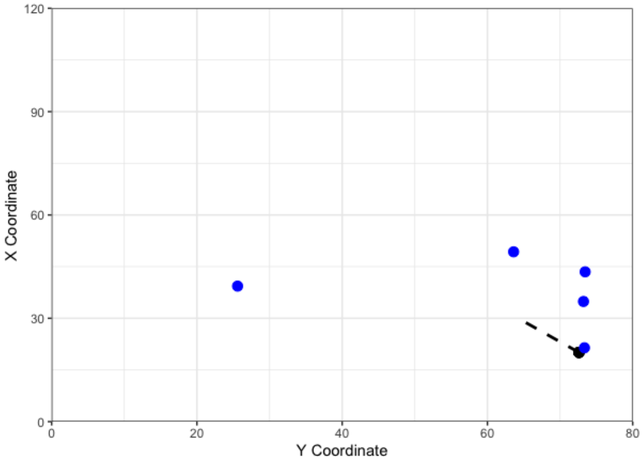


Figure 2: Visualisation of a pass. The black point represents the actor, blue points represent opponents. Dashed line indicates a pass.

Accordingly, for a pass, its starting coordinates, *location.x* and *location.y*, and ending coordinates, *end_location.x* and *end_location.y* are extracted. Furthermore, all the players that are within the frame of the pass are filtered such that only the opponents remain, thereby excluding teammates. This step is taken because this part of the analysis focuses only on the probabilities of the opponents intercepting the ball, rendering the teammates redundant. The locations of the opposing players are also extracted as *ff_location.x* and *ff_location.y*. It is pertinent to briefly describe the manner in which it is determined whether passes are intercepted before proceeding with the next stage of the approach.

A variable *intercepted* is created, which takes the value 1 if a pass is intercepted by the opposing team, and 0 otherwise. As discussed in Section 3, each pass is given a type. Assigning the values 1 and 0 is a straightforward process, as those passes with type "Interception" are labelled 1, and all other passes are labelled 0. The percentage of pass-ids in the dataset with the label 0 is 99.80% ($53,797$ observations labelled 0), indicating that instances of interceptions (1s) are rare. It follows that the data are highly imbalanced.

4.1.2 Predicting Interception Probabilities

To predict the probability of a pass being intercepted, two models are used and their results and performance are compared - the model of choice based on several evaluation criteria will be selected to conduct the analysis. These models are a logistic regression model, and a Random Forest model. The target variable, *intercepted* is a binary outcome that takes the value 1 if the pass is intercepted by the opponent, and 0 otherwise. The nine predictors are the variables listed in Table 1, excluding the *id* variable which serves as an identifier.

4.1.2.1 Logistic Regression

As the target (response) variable is binary, there are several reasons as to why a linear regression model is inappropriate for modelling. Linear models attempt to explain the relationship between a continuous response variable as a linear combination of predictor variables. However, a binary response variable is not continuous. Furthermore, Fritz and Berger (2015) states that it is possible estimates of the response variable that are achieved through least squares are not bounded between 0 and 1 when applying least squares to a binary response variable. As the estimate of the response variable is intended to be interpreted as a probability, said estimates do not provide any useful information.

Generalised Linear Models (GLMs) offer a means of circumventing the issues posed by linear regression techniques on binary data. These are a class of models which enable the modelling of more diverse types of data relative to linear regression models by including a random component, and a link function. Given the values of the predictors, the random component specifies the conditional distribution of the target variable; commonly, this distribution belongs to the exponential family. Faraway (2010) states that the response variable belongs to the exponential family if its probability density function is of the form

$$f\left(y|\theta,\phi\right)=\exp\left(\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(y,\phi\right)\right)\,, \tag{1}$$

where y is a realisation of the random variable, θ is a canonical parameter that represents location, ϕ is a dispersion parameter denoting the scale, and $a\left(\cdot\right), b\left(\cdot\right)$, and $c\left(\cdot\right)$ are specific functions; different members of the exponential family are obtained by defining the specific functions. If $y \sim \mathit{Binomial}\left(n,\pi\right)$, for $a\left(\phi\right)=\frac{1}{n}$, $b\left(\theta\right)=-\log\left(1-\pi\right)$, and $c\left(y,\phi\right)=\log\binom{n}{n\mu}$ where $\mu=\mathbb{E}\left(\frac{y}{n}\right)$, the binomial distribution is obtained.

The link function, $g\left(\cdot\right)$, is responsible for transforming the mean of the target variable, μ, to a linear combination of the predictors. Mathematically, it follows that

$$g\left(\mu\right)=\eta=\alpha+\beta_1 x_1+\ldots\beta_p x_p\,, \tag{2}$$

where η is a linear predictor, expressed as a linear function of the regressors $x_1,\ldots,x_p$ with corresponding coefficients $\beta_1,\ldots\beta_p$.

When a binomial GLM is fitted with a logit link function, a logistic regression is performed - the logistic regression is a special case of a binomial regression. The logit link function is defined as

$$g\left(\mu\right)=\log\frac{\mu}{1-\mu}. \tag{3}$$

From the logit link function, it follows that $0\leq\mu\leq1$. Fox (2015) provides further insights as to how GLMs are estimated and tested.

Thus, to perform the logistic regression, the target variable *intercepted* is defined as y with linear predictor η, where

$$\begin{aligned}\eta=\alpha+&\beta_1\,\mathit{duration}+\beta_2\,\mathit{pass.length}+\beta_4\,\mathit{pass.angle}+\beta_4\,\mathit{location.x}+\beta_5\,\mathit{location.y}+\\&\beta_6\,\mathit{ff\_location.x}+\beta_7\,\mathit{ff\_location.y}+\beta_8\,\mathit{end\_location.x}+\beta_9\,\mathit{end\_location.y}+\epsilon\,,\end{aligned} \tag{4}$$

and α and ϵ are a constant term, and an error term respectively.

To estimate the logistic regression model, group k-fold cross-validation (GCV), with $k=5$ is performed using `caret` (2023). The choices of k, as well as cross-validation instead of repeated cross-validation, are to ensure computational feasibility.

Furthermore, GCV is used due to the structure of the data: as discussed earlier, each unique pass-id consists of multiple rows corresponding to the players that are visible in the frame of the broadcast camera during the game. GCV splits the data in a manner which ensures that no pass-id is contained in both, the training and validation folds. Accordingly, rows from a specific pass-id belong to either the training fold, or the validation fold - but not both. This condition is necessary, for if some rows of a pass-id are placed in the training fold and some in the validation fold, the model is being trained on an incomplete passing scenario. By ensuring that instances of a specific pass-ids are not divided amongst the training and validation folds, the resulting values are more likely to accurately reflect the probability of a pass being intercepted.

To this end, GCV with five folds divides the data into five (roughly) equal sized groups, wherein four of the groups serve as training data, and one serves as the validation data. Each group has a distinct set of pass-ids, and in this iterative procedure, is assigned as the validation set once, such that the model deals with new, unseen data each time. As a result, the model can be evaluated in a more comprehensive manner.

Predictions are made with the logistic regression model on the entire dataset. Hence, for each row of a pass-id, the probability of each player involved in the frame of the pass intercepting said pass is estimated.

Given the class imbalance in the data, a second logistic regression model is estimated using GCV with $k = 5$. However, to address the imbalance in the data, a method of up-sampling is performed using `caret` (2023), where instances of the minority class (1s) are randomly over-sampled with replacement in the folds until they are of the same proportion as instances of the majority class (0s).

Up-sampling is chosen instead of down-sampling, as several studies indicate that based on certain evaluation metrics (such as, but not limited to recall), the former provides more accuracy than the latter (Batista, Prati, & Monard, 2004; García, Sánchez, & Mollineda, 2012). Moreover, there is no loss of information when using up-sampling, as observations are replicated, rather than discarded as is the case with down-sampling. A caveat is that the up-sampling approach may result in over-fitting. Nonetheless, Batista et al. (2004) finds that the performance of random over-sampling is often comparable to increasingly complex over-sampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE). Notably, Batista et al. (2004) additionally indicates that *"class imbalance does not systematically hinder the performance of learning systems."*

Therefore, predictions are made on the original dataset with the second logistic regression model, likewise to the logistic regression model that is trained with the imbalanced, original data. The goal is to examine whether up-sampling the minority instance during the cross-validation procedure yields superior classification and prediction performance. The evaluation procedure is considered at a later stage.

The two logistic regression models are subject to comparison against two Random Forest models to determine which of the two methods is superior overall when dealing with imbalanced and balanced data. The methodology pertaining to the Random Forest is discussed below.

4.1.2.2 Random Forest

Introduced in Breiman (2001), Random Forests are a non-parametric tool used for classification or regression. A Random Forest operates by growing multiple, independent decision trees. Each tree provides a classification, and the Random Forest algorithm chooses the most common classification to be assigned to the data point that is being predicted.

To grow a tree, the algorithm performs bootstrapping. That is, repeated samples of observations and variables are taken from the training data. The algorithm then performs a bagging procedure, which is a term used to refer to bootstrap aggregation. In this step, multiple decision trees are created, wherein each tree is trained on a different bootstrapped sample. The resultant decision trees are therefore less likely to overfit to the data than other tree-based methods which do not perform bagging. Majority voting is performed to obtain a classification, and the probability of the classification for each class is obtained by calculating the proportion of each decision on all the classification trees.

A GCV procedure is implemented to optimise the Random Forest. This optimisation occurs by tuning the *mtry* hyper-parameter. When forming each split in a tree, the Random Forest algorithm randomly considers *mtry* number of variables from the available predictors. Therefore, for each split, the Random Forest considers a different set of predictors - for a classification task such as the one being performed, the default value of *mtry* is $\sqrt{p}$, where p is the number of predictors; in this case, $p = 9$, resulting in a default *mtry* value of 3.

Whilst the Random Forest algorithm contains several hyper-parameters that can be tuned, only *mtry* is considered; Probst, Wright, and Boulesteix (2019) reasons that Random Forests often work reasonably well with their default hyper-parameter values in most applications, thus rendering tuning as unnecessary. Moreover, Wright and Ziegler (2017) finds that computation times when using a Random Forest decrease approximately linearly for lower values of *mtry*, as the algorithm spends a significant amount in selecting candidate variables for the split. As computational feasibility is crucial, a grid-search with $mtry = \{2, 3, 4\}$ is conducted during the GCV procedure to tune the hyper-parameter. The number of trees, *ntree* is set at its default value of 500.

The resulting Random Forest model, with optimal *mtry* is trained on the data and used to make predictions, once again yielding estimates of the probability of each opponent player involved in a pass intercepting said pass for each pass-id in the data. More specifically, for each row in the data, the Random Forest model estimates the probability of said row belonging to the minority class 1, and the probability of it belonging to the majority class 0.

Similarly, to address the class imbalance issue, a second Random Forest model is optimised using a GCV procedure with $k = 5$; *mtry* is tuned in an identical manner as above, with $ntree = 500$; up-sampling is performed within the GCV procedure. The resultant Random Forest model is trained on the data and used to make predictions.

Essentially, two Random Forest models - one which uses original, imbalanced data and one which uses up-sampled data - are trained and tested with three different values of the *mtry* hyper-parameter. Hence, a comparison is facilitated between the performance of the (im)balanced logistic regression model and the (im)balanced Random Forest model, from which one model is selected based on the evaluation criteria described below.

Upon completing these steps, the probability of each opponent player that is visible in the frame of a pass intercepting said pass is known. For instance, in Figure 2, the probability that each opposing player intercepts the pass is known as a result of applying the logistic regression model or the Random Forest model, as these models provide estimates of the probabilities.

4.1.2.3 Model Evaluation

Prior to continuing with the methodology to evaluate the risk of a pass, it is required to evaluate the four models in order to select one to perform the following analysis. The evaluation criteria of choice are: precision, recall, and Brier Scores.

Precision is a measure of the ratio of observations classified as positive by the model and actual positives. Recall - also defined as the True Positive Rate (TPR), or Sensitivity - denotes the percentage of actual positives that are identified as positive by the model.

Saito and Rehmsmeier (2015) demonstrates that precision is capable of extracting differences in the performance of models when dealing with imbalanced data, whereas metrics such as Accuracy (fraction of correct predictions and all predictions) are not able to do so. The precision, valued between $0 - 1$ of a model is easily interpreted as the percentage of correct predictions amongst the positive predictions. Models with a higher value of precision are preferred, as they are more adept at distinguishing between passes that are intercepted or not relative to models with lower precision.

Recall is also measured on a scale between $0 - 1$, and higher values are similarly preferred. It can be interpreted as the percentage of observations that the model is correctly able to identify as belonging to the positive class; in this context, the recall of a model is, from all the passes that are labelled as intercepted, the percentage of these passes the model is able to correctly identify.

The trade-off between precision and recall is well documented, as explained in Buckland and Gey (1994). This trade-off is inevitable when classification performance is consistently better than random. As the analysis deals with imbalanced data, achieving high recall while maintaining a high level of precision is especially challenging. Thus, models which are able to do so are preferred.

The function of the models in Section 4.1.2 is to assign each observation with a probability of belonging to each of the two classes. To evaluate the performance of these models in estimating probabilities, the Brier score is employed. It is calculated as

$$BrierScore = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_i} - y_i\right)^2 \ , \tag{5}$$

where n is the total number of observations in the data, y_i is the observed data, and $\hat{y_i}$ is the prediction of the observation made by the model. Lower values of the Brier score are preferred, as they are indicative of better prediction performance with respect to probabilities. Rufibach (2010) makes several notes with respect to the Brier score; firstly, as defined in equation (5), the Brier score is equal to the Mean Squared Error of prediction. Secondly, a lower Brier score does not necessarily imply better calibration, as the metric is a measure of calibration, in addition to sharpness; the latter is "*the concentration of the predictive distribution.*" To this end, Rufibach (2010) recommends that calibration be addressed when comparing models through the Brier score. However, as the models are not solely evaluated on the basis of their Brier score, but also on the basis of their precision and recall, this issue is not addressed further. To determine the optimal value of *mtry* for the Random Forest models in Section 4.1.2.2, the value in the grid that produces the lowest Brier score is selected as the optimal *mtry*.

To summarise, of the four models, the one with the most favourable precision, recall, and Brier score will be used to predict the probabilities of the (counterfactual) passes being intercepted by their opponents. Having determined how to evaluate and select the final model, the analysis can be continued.

At this stage, the passes are filtered such that only passes that are played under pressure remain, as the purpose of this analysis is to analyse the decision-making for said passes. Therefore, to calculate the risk for all the passes that are in the data is redundant.

4.1.2.4 Obtaining the Probability of Intercepting a Pass

Once the probability of interception is determined for each relevant opponent player for a given pass it is necessary to combine these values into a singular probability of the pass being intercepted. This step is largely based on equation (2) in Burriel and Buldú (2021), which calculates the interception probability, $I\left(init, end\right)$, of a pass starting at its initial location, *init*, and its destination, *end* as

$$I\left(init, end\right) = 1 - \prod_{p \in \mathcal{P}} \left(1 - \rho_{int}\left(p, init, end\right)\right) , \tag{6}$$

where $\mathcal{P}$ defines the set of opponent players that are associated with a pass-id, p is a specific player from this set, and $\rho_{int}\left(p, init, end\right)$ represents the probability of a player p intercepting the pass with starting location *init* and ending location *end*. The calculation assumes that the probabilities of interception are independent of the players.

Consequently, for each pass-id in the data, there is a probability that it is intercepted by the opposition team. It is important to note here that since $I\left(init, end\right)$ is a probability, each value is bounded between the interval $[0, 1]$.

4.1.3 Predicting Possession Probabilities and Measuring the Risk of a Pass

As previously stated, the risk of a pass can be measured by comparing the probability of a pass being intercepted by its opponents, and by its teammates. Having calculated the probability of the former, herein lies the approach with respect to the latter probability.

By replicating the steps to those outlined in the sub-sections above, whilst maintaining that a teammate intends to intercept the ball instead of a rival yields the probability of a pass being intercepted by its teammates. Therefore, passes are now filtered such that only the teammates that are visible in the frame of the pass remain. Figure 3 displays the same pass as Figure 2, but now with the locations of the teammates of the actor as opposed to the opponents, thereby highlighting an example of the situation that is considered in the following analysis.

The target variable *intercepted* takes the value 1 for a successful pass (intercepted by a teammate), and 0 otherwise. The percentage of pass-ids labelled 1 is 83.46% (45,045 observations), implying the existence of class imbalance, albeit a less extreme one than in the interception case.

To estimate the probability of each teammate that is visible in the frame of the pass intercepting the ball, once again logistic regression models and Random Forest models are utilised. Specifically, these models are trained and then used to make predictions on each row of the data. Bearing in mind that each row for a pass-id represents a teammate, probabilities are estimated for each player.

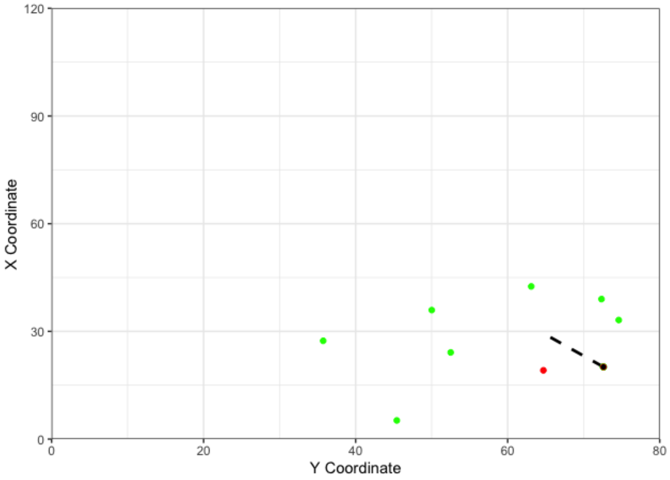


Figure 3: Visualisation of a pass. The black point represents the actor, light green points represent teammates, the red point represents the intended recipient of the pass, determined as the teammate with the closest Euclidean position to the end location of the pass. Dashed line indicates a pass

The same GCV procedure is applied, with $k = 5$ folds to estimate the logistic regression, and optimise the Random Forest models. Hyper-parameter tuning is conducted for the *mtry* parameter in the Random Forest using the same grid-search as in the 'interception' case. After these models are trained, and predictions are made on the data, the GCV procedure is repeated with up-sampling. Having obtained the newly trained models, they are used to once again estimate the probability of each teammate intercepting the ball for a passing situation. Consequently, analogous to the 'interception' case, four models exist in this 'possession' case: (im)balanced logistic regression, and (im)balanced Random Forest.

Following the same steps as in Section 4.1.2 yields the probability of each pass being intercepted in the 'possession' case. Let this probability be denoted by $P\,(init, end)$ defined in the same manner as in equation (6); it holds that these values are bounded between the interval $[0, 1]$. Accordingly, the risk of a pass, R, is defined as $R = I\,(init, end) - P\,(init, end)$. The risk is therefore represented as the probability of a pass being intercepted by its own team subtracted from the probability of the pass being intercepted by the opponent team. It follows that the risk of a pass is bounded between the interval $[-1, 1]$, where a value closer to -1 implies low risk, and values close to 1 imply high risk.

4.2 Creating Counterfactual Passes

Insofar, for a pass that has occurred, there is a methodology to determine its risk, based on its probability of interception by its opponents and teammates. However, this is not yet sufficient to evaluate the decision-making of the actor and team; to do so, it is necessary to know which other passes could the actor have potentially played. The passes that the actor could have played, but did not, are counterfactual passes. Consider Figure 3: in this plot, there are seven other teammates that the actor could have chosen to pass the ball to (denoted by the green points). These seven other passes represent the counterfactual passes; because these passes do not actually occur, some of their characteristics - namely the angle, length, and duration of the pass - are unknown and therefore must be estimated from the given data.

4.2.1 Calculating Pass Length and Pass Angle

For a counterfactual pass scenario such as in Figure 3, the following components are known: the starting coordinates (as the actor is the same), and the locations of each player in the frame of the pass. Obtaining the counterfactual pass length is a relatively straightforward process: assume that for a counterfactual pass, its end location will be the same as the location of the player who receives the ball. It follows that the end location of the counterfactual pass is the *ff_location* of the player who receives the counterfactual pass. The length of the counterfactual pass is obtained as the Euclidean distance between the starting coordinates and ending coordinates of the pass.

Moreover, given the coordinates that specify the starting and ending locations of a counterfactual pass, the counterfactual pass angle is derived as

$$\Delta x = \mathit{ff\_location.x} - \mathit{location.x} \tag{7}$$

$$\Delta y = \mathit{ff\_location.y} - \mathit{location.y} \tag{8}$$

$$\mathit{pass.angle} = \arctan \frac{\Delta y}{\Delta x}. \tag{9}$$

4.2.2 Predicting Duration

The duration of the counterfactual pass, requires a greater amount of inference relative to the other counterfactual variables because it depends on several unknown factors such as, but not limited to, the force and speed with which the ball is received and passed by the actor, weather conditions, and the height of the pass. Thus, this thesis opts to predict the duration of the counterfactual pass, relying once again on the prediction techniques that are employed at several other stages of the methodology.

In Section 3, it is shown that the *duration* variable does not have a linear relationship with a majority of its predictors. Therefore, a Random Forest model is tasked with estimating the value of *duration* for the counterfactual passes. As previously stated, the Random Forest method is non-parametric. Consequently, it does not make any distributional assumptions in its computation.

In order to optimise a Random Forest model to predict the duration of a pass, cross-validation (CV) is opted for with $k = 5$ folds. GCV is not required here, as the data only consist of rows with the original pass and its corresponding information, thereby excluding the rows that represent the additional players in the broadcast camera.

Crucially, the Random Forest model does not perform probabilistic classification to obtain the predicted values of the duration of a counterfactual pass. As *duration* is a continuous variable, not a binary one, Random Forest regression is performed, using the pass variables that are known for a counterfactual pass as predictors. These variables are $pass.angle$, $pass.length$, $location.x$, $location.y$, $end\_location.x$, and $end\_location.y$.

To perform hyper-parameter tuning on *mtry*, the default value of this hyper-parameter is $\frac{p}{3}$ as a Random Forest regression is being performed, where p is the number of predictors, which in this case is 6. Here, the default *mtry* is 2. A grid-search is conducted with $mtry = 1, 2, 3$ within the CV procedure, and the default number of trees, $ntree = 500$, is opted for. The prediction metric used to decide the optimal *mtry* is the *Root Mean Squared Error* (RMSE); values of the hyper-parameter which minimise the RMSE are optimal. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(duration - \widehat{duration}\right)^2}, \tag{10}$$

where n is the number of observations for which predictions are made, and $\widehat{duration}$ is the predicted value.

The final model that has the optimal value of *mtry* is trained on the data and predictions are made, yielding the model estimates of *duration*. To evaluate the accuracy of these predictions, the RMSE metric is used. Lower values of the RMSE are preferred, as they imply that the predictions are closer to the actual values of the *duration*. From here onwards, the duration of a counterfactual pass can be predicted, allowing for the next step of the analysis: obtaining the risk of a counterfactual pass.

4.2.3 Measuring Risk

The risk of a counterfactual pass is estimated in a similar manner as is done previously: the probability of the pass being intercepted by its own team subtracted from the probability of the pass being intercepted by its opponents - therefore, the steps taken here to measure the risk are likewise to those in the case of the original pass. That is, the steps from Section 4.1.2 onwards are repeated: for a pass-id, its pass information is updated with its counterfactual pass length, counterfactual pass angle, counterfactual duration, and counterfactual end-location. Essentially, a new teammate is considered as the recipient of the pass. Then, the probability of this counterfactual pass being intercepted by its opponents is estimated, using the prediction model of choice that is trained in the 'interception' case. This step yields the probability of each opponent intercepting the counterfactual pass. Using equation (6), the probabilities are combined into a singular value, such that the probability of the counterfactual pass being intercepted by the opposing team is acquired.

Similarly, the probability of the counterfactual pass being retained by its team is measured. Given the updated counterfactual pass information, the prediction model of choice trained for the 'possession' case is applied on the data, resulting in the probability of each teammate obtaining the ball. Utilising equation (6) once again provides a probability of the pass being intercepted by its team. Passes are filtered such that the risk is only calculated for those that are played under pressure; that is, the counterfactual risk is only determined for those passes that are originally played under pressure.

The risk of a counterfactual pass is determined in the same way as the risk of the original pass. As discussed in Section 3, for each pass-id, the number of counterfactual passes corresponds to the number of teammates who are visible in the frame of the pass who do not originally receive the pass. Thus, the process described above is repeated with each teammate being assumed as the recipient of the pass, for each pass-id that is considered in the analysis.

Consider the following stylised example presented in Figure 4, which shows the the entirety of the passing situation that is displayed in Figures 2 and 3. Figure 4 shows the original pass, alongside the positions of the teammates and opponents, in addition to the risk of the original and counterfactual passes.

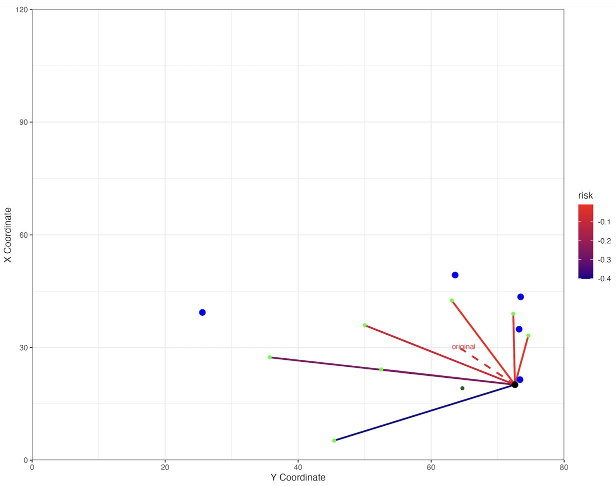


Figure 4: Visualising the risk of a (counterfactual) pass. Blue points represent opponents, green points represent teammates, the black point represents the actor, and the dark green point represents the intended recipient of the pass, determined as the teammate whose location is the smallest Euclidean distance from the end location of the pass. The dashed line is the original pass

Ultimately, this plot represents the culmination of all the methods that have been applied thus far, such that for a pass that occurs under pressure, its risk is defined, plus the risk of all the other passes that could have been played by the actor. Algorithm 1 presents an overview of the required steps to predict the risk of the (counterfactual) passes.

Once the risk of a (counterfactual) pass is ascertained, the next component that is required to assess decision-making is the value of a pass - namely, did the ball move from a less to more valuable part of the pitch as a result of the pass? Did there exist, at the moment that the original pass was made, another passing option that would have added greater value? In the context of this analysis, the value of a pass is regarded as its role in increasing the team's chances of scoring a goal. The next steps are designed to quantify the value of a pass, in order to be able to answer the questions posited above and enrich the depth of analysis.

Algorithm 1 Algorithm to measure the risk of a (counterfactual) pass

```
1: **Initialise:** $all\_passes$, all pass data with $freeze\_frame$ included; $k = 5$
2: Filter passes to only include opponents    ▷ Probability of opponent intercepting the pass
3: Assign $intercepted$ target variable to each pass
4: Train RF and Logistic Reg. using GCV with $k$ folds ▷ group by $id$, optimise $mtry$ for RF; Section 4.1.2
5: Make predictions on data    ▷ Estimate probability of each visible opponent intercepting the pass
6: Repeat Steps 4 and 5 with up-sampling
7: Filter passes to only include teammates    ▷ Probability of teammate intercepting the pass
8: Repeat Steps 4-6    ▷ Estimate probability of each visible teammate intercepting the pass
9: Train RF to predict $duration$ of passes with $k$-fold CV    ▷ Optimise $mtry$ for RF; Section 4.2.2
10: Select models based on evaluation metrics    ▷ Section 4.1.2.3
11: Measure risk of a pass    ▷ Passes where $under\_pressure = TRUE$; Section 4.1.3
12: Update passes with counterfactual values    ▷ Section 4.2
13: Predict probabilities of counterfactual passes being intercepted and retained
14: Calculate risk of counterfactual passes
```

4.3 Valuing Passes

Valuing a pass requires several steps. Firstly, the pitch is divided into identically-sized zones based on the size of the pitch. For StatsBomb data, soccermatics (2019) states that the pitch dimensions are 120 meters long and 80 meters wide; the coordinates of each ball event are recorded accordingly. Given these dimensions, the manner in which the pitch is divided into zones and events are subsequently assigned to zones is similar to the approach used in Bransen (2017), with some adjustments made to reflect the fact that the dimensions of the StatsBomb pitch are bigger than those considered in Bransen (2017). The total number of zones that are created is denoted by n, such that

$$n = \frac{120}{x_{step}} \cdot \frac{80}{y_{step}}\,, \tag{11}$$

where x_{step} and y_{step} are the length and width of the zones, respectively. These values are chosen in a manner that allows for $n = 1600$, yielding $x_{step} = 2.4$ and $y_{step} = 2.5$. It follows that the pitch consists of 50×32 regions.

A total of 1600 zones is desirable because in order to assign a value to each zone on the pitch, the Expected Possession Value (EPV) is utilised. Fernández, Bornn, and Cervone (2019) defines EPV as the expected outcome of a possession, and its value represents the probability of a possession sequence ending in a goal, given the current location of the ball. Through Friends of Tracking Data (2021), an EPV matrix based on historical data from several leagues and seasons is obtained. The matrix dimensions consist of 32 rows and 50 columns, thus, it holds that there are 1600 elements in the matrix, each corresponding to a zone on the pitch. Each element corresponds to the value of a particular zone; as previously discussed, these values are determined by analysing possession outcomes and events of multiple games across several years. The value of a pass is then obtained by computing the difference in EPV of the start-zone and end-zone of a pass. The difference, ΔEPV, is denoted as the *Zone Possession Value* (ZPV).

To obtain the ZPV for the original pass corresponding a pass-id, as well as its counterfactual passes, it is required to assign a start-zone and end-zone to the (counterfactual) pass. In doing so, it is important to consider whether the player under consideration is the actor, or a teammate. For an original pass, its start location is given by the variables $location.x$ and $location.y$, whereas its end location is given by the variables $end\_location.x$ and $end\_location.y$. Moreover, Section 4.2 states that for a counterfactual pass, whilst the start location remains the same, its end location is assumed to be that of the counterfactual recipient. Therefore, the end location of the counterfactual pass is given by the variables $ff\_location.x$ and $ff\_location.y$ of the teammate that is assumed to be the new recipient of the pass.

To this end, the original pass is assigned to a start-zone and end-zone, z_{start} and z_{end}, respectively as follows

$$\begin{aligned}
x_{adj} &= \left(\lfloor \frac{location.x}{x_{step}} \rfloor - 1 \right) \cdot \frac{80}{y_{step}} \\
y_{adj} &= \lfloor \frac{location.y}{y_{step}} \rfloor \\
x_{end} &= \left(\lfloor \frac{end\_location.x}{x_{step}} \rfloor - 1 \right) \cdot \frac{80}{y_{step}} \\
y_{end} &= \lfloor \frac{end\_location.y}{y_{step}} \rfloor \\
z_{start} &= \begin{cases} y_{adj} & \text{if } x_{adj} < 0 \\ x_{adj} + y_{adj} & \text{otherwise} \end{cases} \\
z_{end} &= \begin{cases} y_{end} & \text{if } x_{adj} < 0 \\ x_{end} + y_{end} & \text{otherwise}\,, \end{cases}
\end{aligned} \tag{12}$$

where 80 represents the length of the pitch, and $\lfloor \cdot \rfloor$ denotes the floor function. The counterfactual passes are assigned to zones using the same approach, with the only difference being that *end_location.x* and *end_location.y* are replaced by *ff_location.x* and *ff_location.y*. It follows that upon performing this step, each (counterfactual) pass has a start-zone and end-zone with a corresponding EPV value. As a result, subtracting the end-zone EPV from the start-zone EPV provides the ZPV for each (counterfactual) pass. That is, the value of each original pass is known, as well as the value of all the other passes that could have been played instead.

Currently, there exists a methodology to evaluate the risk of a pass based on its probability of being intercepted by its opponents or teammates as detailed in Section 4.1. Additionally, there is a process to measure the value of a pass, given in Section 4.3. The concluding step of the methodology is to quantify and subsequently evaluate decision-making.

4.4 Quantifying Decision-Making

For a given pass, it is now known what the risk is, as well as the value (quantified as the ZPV). For all other possible passes in the frame, the above components are also known (counterfactual risk and counterfactual ZPV). Consider the stylised example in Figure 4. To determine whether the selected pass is the best possible option based on the premises of risk and ZPV or whether a counterfactual pass exists that proves more valuable in the context of these metrics, two measures are considered: a *risk Decision Parameter* (*rDP*), and a *Risk-Value Efficiency* (*RVE)* metric.

4.4.1 Risk Decision Parameter

The *rDP* is defined as the percentage of passes made by a team to the player with the lowest risk - in other words, this parameter looks at the number of occasions on which the actor chose to play the pass least likely to be intercepted by the opponent, and most likely to be intercepted by the teammates. Therefore, it follows that the *rDP* aids in quantifying the risk-aversion of a team. The *rDP* of a team over the entirety of the tournament can be calculated as

$$rDP = \frac{\text{number of passes made to player with lowest risk}}{\text{total number of passes made}}. \tag{13}$$

The *rDP* can be further investigated on a stage-by-stage basis or for a specific match to analyse passing behaviour at various moments of the tournament.

4.4.2 Risk-Value Efficiency Parameter

Likewise, to study the manner in which teams balance the risk and value they assume when making a pass, a *value Decision Parameter* (vDP) is introduced as an intermediary measure, where

$$vDP = \frac{\text{number of passes made to player with highest ZPV}}{\text{total number of passes made}}. \tag{14}$$

From all the passes made by a team during the tournament, the vDP looks at how many of these passes are made to the player with the highest ZPV. Therefore, the vDP measures the percentage of passes by a team that move the ball to a more valuable area (zone) of the pitch in terms of the EPV of the zone. Then, the RVE of a team over the entirety of the tournament is defined as

$$RVE = \frac{\text{total } vDP \text{ of team}}{\text{total } rDP \text{ of team}}. \tag{15}$$

The RVE of a team can be interpreted in the following manner: a higher RVE value indicates that the team generates more value relative to the risk of their passes. Consequently, this parameter augments the insights provided by the rDP by helping determine a greater extent of risk-aversion; that is, for teams that are indeed risk-averse, is this tendency also reflected in a lack of willingness to move the ball to more valuable areas of the pitch?

4.4.3 Mann-Whitney U Test

In order to test whether the values of the decision parameters for teams are significantly different from each other, certain groups of teams and their associated rDP values are compared. Specifically, the eight best teams (those that qualify for the Quarter-Finals and beyond) are compared to the other 24 teams to investigate whether there are irrefutable differences in the risk-behaviour of these teams. As there are only 32 teams in the tournament and each team has a corresponding rDP value, the sample size is not sufficient to draw conclusions regarding the distribution of the values corresponding to the decision parameter. Therefore, a non-parametric test, namely the Mann-Whitney U test (also known as the Wilcoxon Rank-Sum test) is applied. The null hypothesis of the test is that there is no significant difference in the decision parameter value between groups, and it is rejected for a p-value < 0.05. Described in McKnight and Najab (2010), this test "*assesses whether two independently sampled groups differ on a single, continuous variable.*"

5 Results

5.1 Model Selection

There are two distinct cases that must be evaluated in order to obtain models that will help answer the research question. These cases are as follows: models for predicting (i) interception probabilities (by opponents); and (ii) possession probabilities (by teammates). Each case contains four models - two Random Forest models, and two logistic regression models. One Random Forest model and logistic regression uses the original, imbalanced data in their respective cross-validation procedures, whereas the other Random Forest model and logistic regression use up-sampled data in their cross-validation procedure.

These models are compared and evaluated based on the steps which are outlined in Section 4.1.2.3. The results are obtained using a MacBook Air M1 2020, 16GB RAM, running RStudio Version $2023.03.0 + 386$.

5.1.1 Models for Predicting Interception Probabilities

Here, one of four models will be selected based on the evaluation procedure in order to use for the purpose of predicting the probability of the original pass and counterfactual pass being intercepted by its opponents. The averages (and standard deviations) of the precision, recall, and Brier Score from performing the cross-validation procedure are presented in Table 4.

For both Random Forest models, the optimal value of *mtry* following a grid-search during 5-fold GCV is 2, based on the Brier scores presented in Table 3; as stated earlier, the performance metric used to select the hyper-parameter value is the Brier score. An *mtry* of 2 produces the lowest Brier score across all models. Additional cross-validation metrics for each value of the *mtry* grid are also shown in Table 3

From the results presented in Table 4, it is evident that both Random Forest models largely outperform the logistic regression models across the three evaluation criteria during cross-validation. The exception lies in the average Brier score, wherein the logistic regression model achieves a better performance than the up-sampled Random Forest although the difference in scores is negligible.

The logistic regression models perform poorly in terms of average precision and average recall relative to the Random Forest models; their scores indicate that the logistic regression models do not possess any classification ability. However, their average Brier scores suggest that the probabilistic predictions made by the logistic regression model are accurate; providing up-sampled data during cross-validation serves to worsen the quality of predictions, as evidenced by the larger average Brier score of the up-sampled logistic regression.

Table 3: Cross-validation metrics of the Random Forest (RF) models used to predict interception probabilities for different values of *mtry*.

	Evaluation Criteria		
Model	Precision (SD)	Recall (SD)	Brier Score (SD)
		$mtry = 2$	
RF	0.00239 (0.000865)	0.125 (0.0364)	0.00187 (0.000293)
RF (up-sampled)	0.00362 (0.00272)	0.120 (0.0665)	0.00190 (0.000308)
		$mtry = 3$	
RF	0.0028 (0.000935)	0.0975 (0.0412)	0.00189 (0.000292)
RF (up-sampled)	0.00439 (0.00293)	0.118 (0.0736)	0.00197 (0.000319)
		$mtry = 4$	
RF	0.00284 (0.00130)	0.114 (0.0452)	0.00190 (0.000286)
RF (up-sampled)	0.00402 (0.00435)	0.0848 (0.0964)	0.00197 (0.000319)

As discussed in Section 4.1.2.3, achieving high precision and recall is a challenge with imbalanced data - such is the case for the models even after up-sampling. Given the similarity of the Random Forest models in terms of their performance, the Random Forest model that uses the original, imbalanced data in its cross-validation is preferred to predict interception probabilities, as it has a higher average recall and lower average Brier score. This Random Forest will be further employed to obtain the interception probabilities of the (counterfactual) passes.

Training the selected Random Forest model and making predictions on the original passes produces the following results: precision = 1.000, recall = 1.000, and Brier score = 0.0000141. These results indicate that the model is over-fitting to the data. Moreover, as the counterfactual passes do not actually take place, predictions made for these passes cannot be entirely assessed. Whilst the tendency of the model to over-fit affects the reliability of these estimates, its Brier scores in both, cross-validation and prediction suggest that the model can make appropriate probabilistic predictions for the purpose of this analysis.

Table 4: Cross-validation metrics of the Random Forest (RF) and logistic regression models (Logistic Reg.) used to predict interception probabilities, averaged across five-folds. Standard deviation is reported as SD.

	Evaluation Criteria		
Model	Avg. Precision (SD)	Avg. Recall (SD)	Avg. Brier Score (SD)
RF	0.00239 (0.000865)	0.125 (0.0364)	0.00187 (0.000293)
RF (up-sampled)	0.00362 (0.00272)	0.120 (0.0665)	0.00190 (0.000308)
Logistic Reg.	0.000 (0.000)	0.000 (0.000)	0.00189 (0.000290)
Logistic Reg. (up-sampled)	0.000 (0.000)	0.000 (0.000)	0.170 (0.00450)

5.1.2 Models for Predicting Possession Probabilities

As discussed earlier, the steps taken in Section 4.1.2 are repeated with teammate data rather than opponent data, creating a similar situation in terms of evaluation to the one above. Four models are to be evaluated: two logistic regression models, one of which uses up-sampled data in its cross-validation procedure, and two Random Forest models wherein one also uses up-sampled data in its cross-validation procedure.

After tuning the *mtry* hyper-parameter, the resulting optimal value is 2 for both Random Forest models. Table 5 showcases the evaluation scores of the four models, where for $mtry = 2$, the lowest Brier scores are achieved. Table 6 presents the results across the five folds during the cross-validation procedure.

Once again, the superiority of the Random Forest models is apparent, as they outperform the logistic regression variants in average precision, recall, and Brier score. The up-sampled Random Forest does achieve more preferable scores to a minor extent; as the difference in scores is not vast, the Random Forest model which does not rely on up-sampling is preferred. Hence, the Random Forest that uses the original, imbalanced data in its cross-validation is selected to estimate the probabilities of the original and counterfactual passes being intercepted by their teammates, as it has a lower Brier score than the logistic regression and up-sampled Random Forest whilst achieving a higher, or a similar level of precision and recall respectively.

Table 5: Cross-validation metrics of the Random Forest (RF) models used to predict possession probabilities for different values of *mtry*.

		Evaluation Criteria	
Model	Precision (SD)	Recall (SD)	Brier Score (SD)
		$mtry = 2$	
RF	0.00896 (0.0200)	0.0000102 (0.0000229)	0.0732 (0.00166)
RF (up-sampled)	0.0229 (0.0511)	0.0000134 (0.0000299)	0.0742 (0.00172)
		$mtry = 3$	
RF	0.00675 (0.00937)	0.0000136 (0.0000186)	0.0739 (0.00163)
RF (up-sampled)	0.0227 (0.0316)	0.0000436 (0.0000622)	0.0748 (0.00173)
		$mtry = 4$	
RF	0.00282 (0.00630)	0.00000669 (0.000150)	0.0739 (0.00163)
RF (up-sampled)	0.0165 (0.0193)	0.0000406 (0.0000515)	0.0751 (0.00178)

Table 6: Evaluation metrics of the models used to predict possession probabilities.

		Evaluation Criteria	
Model	Avg. Precision (SD)	Avg. Recall (SD)	Avg. Brier Score (SD)
RF	0.00900 (0.0200)	0.0000102 (0.0000287)	0.0732 (0.00166)
RF (up-sampled)	0.0229 (0.0511)	0.0000134 (0.0000299)	0.0742 (0.00172)
Logistic Reg.	0.000 (0.000)	0.000 (0.000)	0.108 (0.00159)
Logistic Reg. (up-sampled)	0.000 (0.000)	0.000 (0.000)	0.178 (0.000971)

Training the selected Random Forest model and making predictions on the data gives the following results: precision = 1.000, recall = 1.000, and Brier score = 0.000662. Likewise to the ‘interception’ case, over-fitting is present; using a similar line of reasoning, the model makes reasonable probabilistic predictions, therefore, it is used to predict the probabilities of the (counterfactual) passes.

Overall, to predict interception probabilities of (counterfactual) passes, and to predict possession probabilities of (counterfactual) passes, the corresponding Random Forest models which use the original, imbalanced data during cross-validation are selected. Ultimately, an answer to the sub-question that is presented in Section 2, “*does a logistic regression model offer better prediction* *performance than a Random Forest model on imbalanced data or balanced data?*" is found. It holds that for predicting both, interception and possession probabilities, the Random Forest that uses the original. imbalanced data is preferred, based on metrics such as precision, recall, and Brier score.

5.1.3 Random Forest for Predicting Duration

Having determined which models are used to predict the interception and possession probabilities, the Random Forest model that is developed to predict the duration of the counterfactual passes is considered. Performing 5-fold cross-validation with a grid-search yields an optimal *mtry* value of 2, based on the RMSE scores in Table 7.

Across the five folds, the average RMSE is 0.440, with a standard deviation of 0.0186. After training the optimised model and using it to make predictions on the data, an RMSE of 0.201 is obtained.

Table 7: Cross-validation metrics of the Random Forest model used to predict *duration* for different values of *mtry*.

mtry	**Root Mean Squared Error (SD)**
1	0.442 (0.0183)
2	0.440 (0.0186)
3	0.442 (0.0189)

5.2 Pass Evaluation

Herein lie the results of applying the analytical framework that is constructed, applied, and evaluated in Sections 4 and 5.1.

Each pass that is considered in the analysis is compared to its counterfactuals to quantify the passing-behaviour of teams at different stages of the tournament. Prior to delving into the discussion of the results, it is pertinent to take into account certain things. Firstly, teams with only three games played are those that did not proceed from the Group Stage of the tournament; four games played - did not proceed from the Round of 16; and five games played - did not progress beyond the Quarter-Finals. The two teams that participated in the Final - Argentina and France - played seven games each, whereas the teams that were eliminated in the Semi-Finals, Croatia and Morocco, competed in the 3$^{\text{rd}}$ Place Final, likewise bringing their tally up to seven for the number of games played in the tournament.

Second, and interestingly, Croatia and Morocco played against each other on two separate occasions in the tournament: once in the Group Stage, and once in the aforementioned 3$^{\text{rd}}$ Place Final. Thus, their passing behaviour against each other at two distinct stages of the tournament can be closely inspected, something which cannot be done for any other pair of teams in the tournament. To quantify and evaluate the passing behaviour, the relevant parameters which are introduced in Section 4 are considered below.

5.2.1 Risk Decision Parameter

Recall that the *rDP* measures the number of passes made by the actor to a player with the lowest risk when under pressure - that is, on how many occasions does the actor make a pass that is least likely to be intercepted by the opponent and most likely to be intercepted by a teammate, whilst an opponent player is within a certain proximity of the ball prior to it being passed.

Figure 5 provides an overview of the *rDP* value for each team, grouped by the number of matches played in the tournament. Certain teams stand out, those being Spain (ESP) and Serbia (SRB), who achieve the highest and lowest *rDP* values of all teams in the tournament, respectively.

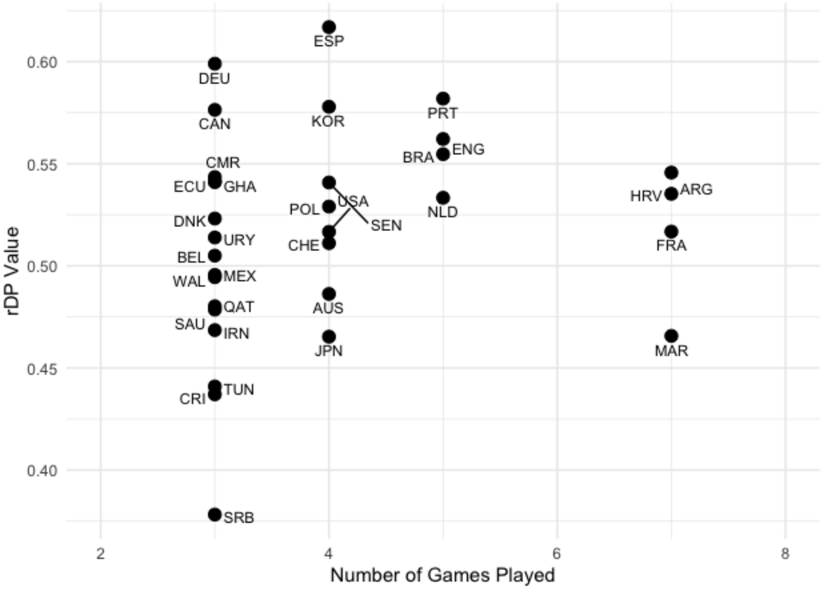


Figure 5: *rDP* values of teams across the tournament

Note: Country names correspond to their iso3c codes obtained from the `countrycode` (2023) package in R. A full list of the names and codes can be found in Appendix A.

Of the 355 passes that ESP play under pressure across their games, 219 are the least risky possible passes, corresponding to a *rDP* value of approximately 0.617. The interpretation is as follows: nearly 62% of passes made under pressure by ESP during the tournament are to the player with the least risk. On the other hand, 59 of the 156 passes that SRB face under pressure are made to the least risky option, yielding an *rDP* of approximately 0.378. 21 teams achieve *rDP* values equal to or higher than 0.50 in the Group Stage, indicating that at least more than half the time, these teams pass to the player representing the safest option; 14 of these 21 teams are eliminated within the first two rounds of the tournament. The results potentially allude to highly conservative teams failing to achieve success in the tournament. Figure 6 plots the *rDP* values of the Quarter-Finalists in each round up until the Quarter-Finals. Notably, in one of the matches played in the Quarter-finals, the winning team, FRA, have a higher *rDP* value than their opponents, ENG. Whereas, in the other three matches, the winners (ARG, HRV, MAR) have lower *rDP* values than their opponents (NLD, BRA, PRT).

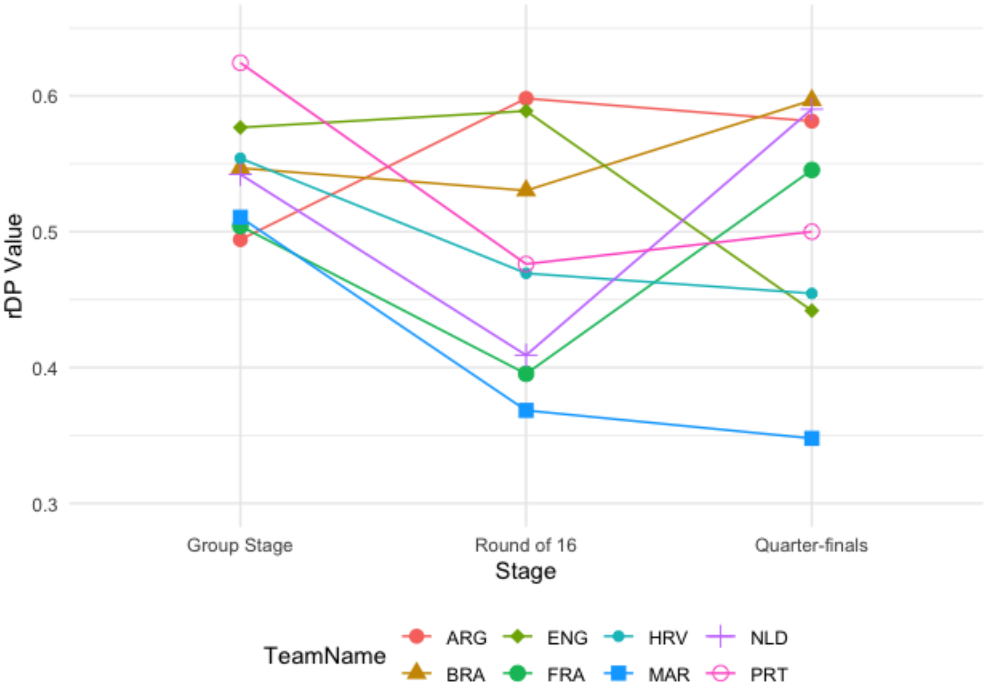


Figure 6: *rDP* by tournament stage of teams that qualified for the Quarter-Finals

On closer inspection, when excluding the five matches where the winner is decided by a penalty shootout and the ten games which end in a draw, the winning team has a higher and lower *rDP* value than their opponent in 23 and 26 games respectively. There are no matches in which the winning team has an *rDP* value which is equal to that of the losing team. It follows that, for games in which the winner is decided through only periods of open-play, teams that are more risk-seeking in passes under pressure tend to win more often than teams that do not exhibit a similar propensity for risk-seeking. Such a result may point towards risk-seeking teams possessing quicker and better decision-making under pressure than their opponents, resulting in greater success.

For the five games that are decided by a penalty shootout, the winner possesses a lower *rDP* value than their opponent. Such a result possibly implies a relationship between risk-seeking behaviour under pressure, and penalty-shootout performance which, likewise to passes under pressure, is a situation in a football match that is characterised by elevated mental pressure. It is possibly the case that this risk-seeking behaviour is a proxy for a measure of confidence, evidenced by these teams' success in the penalty shootout over their less confident opponents.

The corresponding games are ARG-NLD, ARG-FRA, HRV-JAP, HRV-BRA, and MAR-ESP. Figure 5 illustrates that of the winning teams, MAR have a lower *rDP* value over the entire tournament than their opponents ESP. Another winning team, HRV, has a lower *rDP* over the tournament than BRA, but this does not hold for their other opponents JPN, and neither does this result hold for ARG, who have a higher *rDP* than both NLD and FRA. Therefore, while these teams are more risk-seeking during that specific game than their opponent, they are not necessarily more risk-seeking throughout the tournament. Thus, while demonstrating increased risk-seeking behaviour for passes under pressure relative to their opponent may provide an indication of penalty shootout success, it cannot be concluded with certainty that this is the case.

To investigate the the *rDP* further, consider the following five groups: teams that do not progress beyond the (i) Group Stage; (ii) Round of 16; (iii) Quarter-Finals; (iv) Semi-Finals; and (v) Final. Table 8 contains the average *rDP* value of these teams across different stages of the tournament. For instance, the eight teams that progress from the Group Stage but are eliminated in the Round of 16 have an average *rDP* value of 0.524 and 0.544 in those respective stages.

The eight teams that progress to the Quarter-finals and beyond show a decrease in their average *rDP* values upon progressing from the Group Stage to the Round of 16 - that is, they become more risk seeking when passes under pressure are concerned. However, there is no apparent trend in average *rDP* values across stages, as shown by the last row of Table 8. Thus, there is no overall behaviour with respect to risk-taking by teams that can be successfully encapsulated by the *rDP*.

To obtain a better comprehension of teams' attitudes towards risk as they progress further in the tournament, two specific games are inspected closely: as noted previously, MAR and HRV encounter each other twice during the tournament. Their first meeting, which occurred in the Group

Stage, ended in a draw. The *rDP* values corresponding to HRV and MAR in that game are 0.563 and 0.368 respectively. Their next encounter, the match that sees them compete for 3$^{\text{rd}}$ place in the tournament, their *rDP* values are 0.521 and 0.510 respectively. This decrease in the *rDP* corresponding to HRV would seem to indicate that whilst they play with greater caution at the start of the tournament, this is not the case when there is more to lose in terms of a tangible reward (a potential group stage elimination, versus placing third or fourth in the tournament). Importantly, HRV emerged as the winner of their second match-up against MAR.

Similarly, the average *rDP* value achieved by ARG and FRA in the first four stages of the tournament is 0.570 and 0.503 respectively; in the Final, their *rDP* values are 0.444 and 0.606. The deviation in *rDP* from the first four stages to the Finals provides additional evidence of successful teams being cautious with their passing decisions the further they progress, and once there is nothing else on the line, nothing else to play for (final game of the tournament), teams are more willing to take risks with their passing under pressure.

Table 8: Average *rDP* values for teams that are eliminated in each stage of the tournament.

	Average rDP Value				
Last Stage Played by Team (number of teams)	Group Stage	RO16	Quarter-Finals	Semi-Finals	Final
Group Stage (16)	0.501	-	-	-	-
Round of 16 (8)	0.524	0.544	-	-	-
Quarter-Finals (4)	0.573	0.501	0.532	-	-
Semi-Finals (2)	0.532	0.419	0.401	0.564	-
Final (2)	0.499	0.497	0.563	0.587	0.525
All Possible Teams	**0.517**	**0.512**	**0.507**	**0.576**	**0.525**

Note. The average *rDP* value of the 3$^{\text{rd}}$ Place Final is 0.516, played by two teams. The last row provides the average *rDP* values of all the teams that compete in that stage.

Such findings corroborate the research conducted by Mundstock et al. (2021) and Riedl et al.

(2015) that is discussed in Section 2. As demonstrated by the greater extent of risk-taking in the last games of the tournament, successful teams seemingly choose to wait until the last stages of a game (in this case, a tournament) before adapting to a more aggressive approach. This result illustrates an important facet of Prospect Theory described in the foregoing literature: that losses are valued more than gains. On average, teams become risk-averse over the duration of the tournament - they value being able to remain in the tournament to the extent that a safer approach is preferred to a riskier one - until being risk-averse has no consequence apart from being detrimental to their success (winning the final game of the tournament).

Given the average *rDP* values of each group across the stages of the tournament that they compete in, a Mann-Whitney U Test is conducted. The goal is to determine whether the average *rDP* values exhibit statistically significant differences between stages of the tournament. The null hypothesis of the test is that there is no significant difference in average *rDP* values for a stage between teams eliminated in subsequent rounds of the tournament. The results are given in Table 9. This hypothesis is rejected for a p-value less than 0.05. The first value of the first column can be interpreted as follows: for teams that are eliminated in the Group Stage and the Round of 16, there is no significant difference in their average *rDP* values in the Group Stage (0.501 and 0.524, see Table 8).

Table 9: p-values of the Mann-Whitney U Test

	***p*-value**				
Stage of Elimination	Group Stage	RO16	Quarter-Finals	Semi-Finals	Final
Group Stage vs. RO16	1.000	-	-	-	-
RO16 vs. Quarter-Finals	1.000	1.000	-	-	-
Quarter-Final vs. Semi-Finals	1.000	1.000	1.000	-	-
Semi-Finals vs. Final	1.000	1.000	1.000	1.000	-

Based on the results in Table 9, there is no evidence to suggest that the average *rDP* values differ significantly between the specified stages of elimination.

Additional comparisons are performed between the average *rDP* values in the Group Stage of the teams ranked $1-8$ and $9-24$ in the tournament and in the Round of 16 of the teams ranked $1-8$ and $9-16$, to investigate trends in *rDP* values. The results are displayed in Table 10. The null hypothesis, that there is no significant difference in average *rDP* values in the Group Stage between the two groups of teams is not rejected. The same conclusion follows for the null hypothesis that there is no significant difference in average *rDP* values in the Round of 16 between the two groups of teams.

Table 10: Results of the Mann-Whitney U Test performed on groups of teams

	Group Stage		RO16	
	Group 1 (Teams ranked 1-8)	Group 2 (Teams ranked 9-24)	Group 1 (Teams ranked 1-8)	Group 2 (Teams ranked 9-16)
Average *rDP*	0.544	0.509	0.479	0.544
***p*-value**	0.09		0.207	

Ultimately, teams that progress from one round to another do not show differences in risk-seeking behaviour relative to teams that do not progress, as quantified by the *rDP*. Thus, while the *rDP* captures some effect of risk on winningness, it is not statistically significant in explaining the risk assumed by teams in different stages of the tournament. Therefore, it cannot be concluded with certainty that the risk assumed by teams with passes under pressure is linked to their success in the tournament, or the sole determinant of success at the very least. Nevertheless, the conducted analysis provides an answer to the following sub-question: "*how does the risk-seeking behaviour of teams change at different stages of the World Cup?*"

Specifically, in behaviour that is consistent with Prospect Theory, on average teams show a propensity for risk-aversion as they proceed to subsequent stages of the World Cup, except for in the (pen)ultimate game. Some variation in the risk assumed by teams can be explained by the quality of their opponent. For instance, teams may choose to perform riskier passes against opponents who are perceived to be inferior; future research can investigate the impact of implementing betting-odds as an indicator of which team is favoured to win, and use that to weight the amount of risk that a pass contributes to the *rDP* value of a team.

5.2.2 Risk-Value Efficiency

Recall that the *RVE* is the ratio of the *vDP* and *rDP* of a team. Figure 7 shows the *RVE* and number of games played for each team in the tournament. The largest *RVE* value, 0.424, is generated by SRB in their three games. For completeness, the *vDP* and *rDP* values are 0.160 and 0.378 respectively, yielding the stated *RVE* value. The *vDP* value indicates that of all the passes made by SRB in the tournament, approximately 16% are made to the player with the highest ZPV; likewise, the *rDP* shows that nearly 38% of the total passes are made to the player with the lowest risk.

Note that there are no teams with an *rDP* value of 0 over the entirety of the tournament, thereby ensuring that the *RVE* exists for every team. Moreover, every team has an *RVE* value below 1, indicating that *rDP* dominates the *vDP*; that is, each team prioritises making safer passes at the expense of passing to more valuable zones on the pitch.

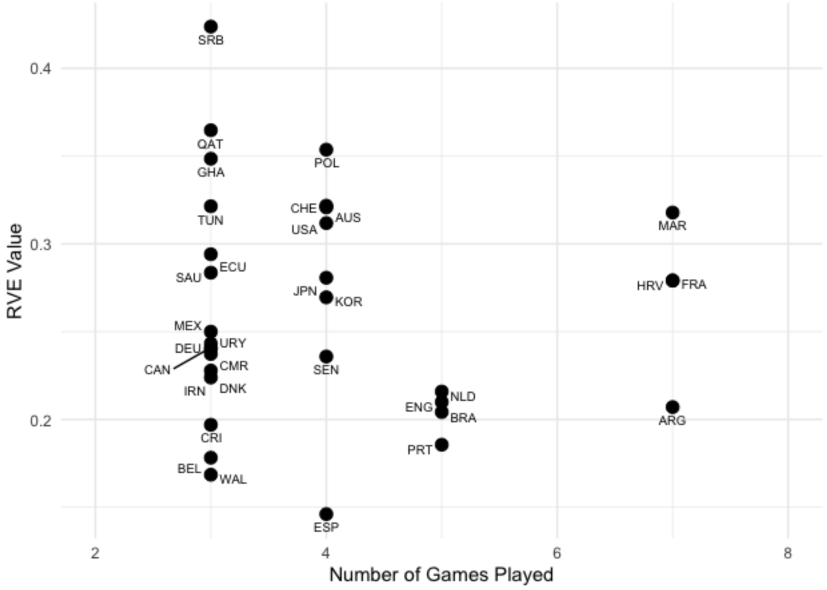


Figure 7: *RVE* of teams across the tournament

For a more holistic evaluation, consider Figure 8, which plots the *rDP* and *RVE* values of each team. The dashed lines that divide the plot into quadrants are the median values of the relevant quantification measures. The median is utilised as it is a better measure of central tendency than the mean, in that it is a more robust method for asymmetric data, and less sensitive to outliers. The plot indicates a negative relationship between *rDP* and *RVE*, suggesting that as a team makes increasingly safe passes under pressure, they pass the ball to less valuable zones on the pitch, ceteris paribus. That is, the tendency to play safe passes is also reflected in a lack of willingness to move the ball to valuable areas of the pitch.

The resultant quadrants display distinct qualities with respect to risk and value: teams in the 'Low Risk, Efficient Passes' quadrant are those with a higher-than-median *rDP* and *RVE*. Accordingly, relative to the other teams in the tournament, the teams in this quadrant prefer to take fewer risks in their passing under pressure, but tend to be 'efficient' in these passes; that is, they attempt to pass more often to players that are in a more valuable zone of the pitch than the actor.

On the other hand, teams in the bottom-right quadrant are those with a higher-than-median *rDP* but lower-than-median *RVE*. Such a position suggests that these teams are less risk-seeking and 'inefficient' in their passes relative to other participants, as they are not moving the ball to valuable areas on the pitch as often as the other teams with their passes. Such teams are therefore conservative in their behaviour with respect to passes under pressure, whilst teams in the 'High Risk, Inefficient Passes' quadrant look to play riskier passes that do not contribute much in moving the ball to valuable zones of the pitch. Teams in this particular quadrant possibly exhibit poor decision-making under pressure, as they assume a greater risk for a lesser payoff more often when compared to other teams.

The 'High Risk, Efficient Passes' quadrant contains those teams that assume less-than-median risk as measured by the *rDP*, and higher-than-median *RVE*. These teams look to make riskier passes with greater frequency than other teams whilst under pressure, and these passes are located to players in valuable zones more often than not. Whether being in this quadrant has any effect on success in the tournament is an important insight. Each team is given a rank from $1 - 32$ based on where they finish in the tournament (see Miles, 2022). To this end, Table 11 records the average finish of the teams in each quadrant.

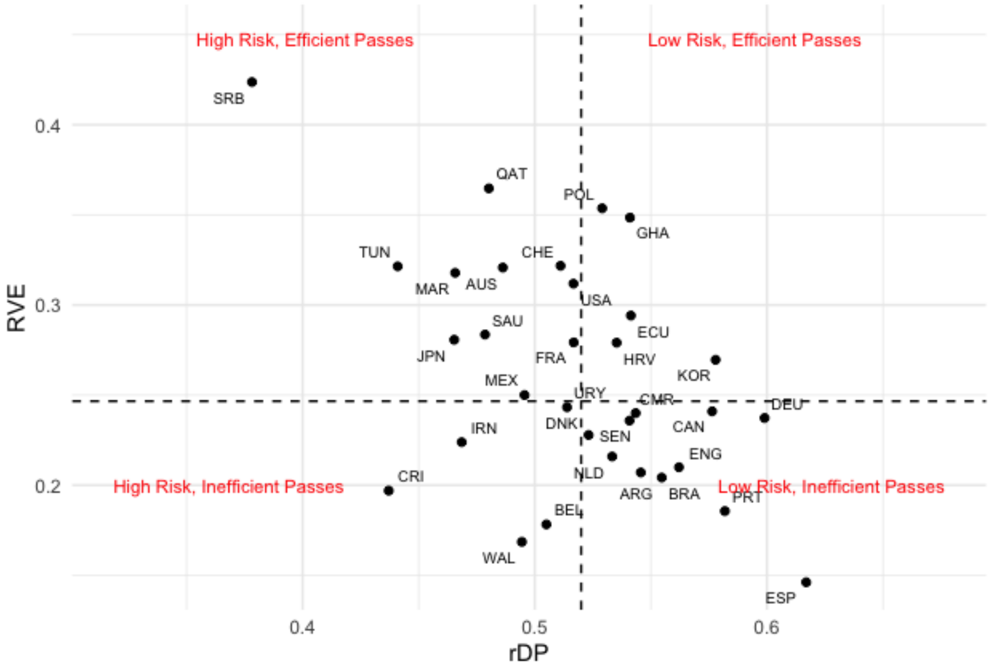


Figure 8: Scatterplot of *rDP* and *RVE*

Teams in the ‘Low Risk, Inefficient Passes’ quadrant have an average rank that is better than the teams belonging to the other three quadrants, implying that teams that are less risk-seeking and do not look to find valuable passes under pressure are better off in terms of success.

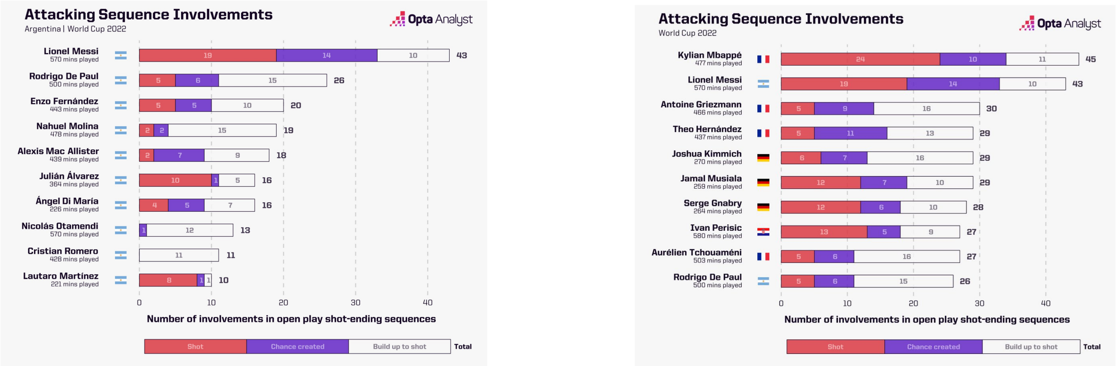
Table 11: Average rank of teams for each quadrant in Figure 8.

Quadrant (number of teams)	**Average Rank**
High Risk, Efficient Passes (11)	16.455
High Risk, Inefficient Passes (5)	26.800
Low Risk, Efficient Passes (5)	15.200
Low Risk, Inefficient Passes (11)	12.455

Conversely, teams situated in the ‘High Risk, Inefficient Passes’ quadrant have the worst average rank of all teams by a considerable amount, suggesting that assuming greater risk in an unproductive manner is not necessarily a viable strategy. In fact, such teams are potentially better off assuming less risk, even if the passes they make are still inefficient.

One possible explanation of the fact that the teams situated in the ‘Low Risk, Inefficient Passes’ quadrant enjoy greater rankings on average, is the individual quality of players possessed by these teams. This factor possibly reduces the propensity of the team to take risks, as they look to find these players over other teammates even when they are not in valuable zones of the pitch, as these individuals are highly capable of adding value from said zones. To this end, Figures 9a and 9b provide support for this claim, as they show that Lionel Messi (ARG) has the most attacking sequence involvements of the ARG players during the tournament, and the second-most overall; only one other player from ARG is found in the latter list. It follows that throughout their games, ARG placed a reliance on Messi and possibly sought to provide him with the ball more often, offering an insight with respect to their *RVE* profile.

Similarly, three players from DEU are present in Figure 9b, which is notable as these players are amongst those with the most attacking sequence involvements despite having played only three games; every other player on the list played seven. This finding is again potentially indicative of the fact that DEU finds these players more often in less risky and less valuable positions, from which said players can create value through their abilities. Other notable teams in this quadrant include Quarter-Finalists BRA and ENG; these teams may possibly also express a reliance on their star players to create value from low risk situations.



(a) ARG players

(b) Ten players with most involvements

Figure 9: Number of attacking sequence involvements during the tournament

Note: Figure 9a is obtained from Furniss, Sisneros, and Manuel (2022).

Betting odds prior to the start of the tournament are provided in Piacenti (2022) and used to further complement the findings in Table 11. Odds with a '+' sign and '−′ sign indicate an underdog and favourite, respectively. These odds (of teams qualifying from their respective groups, and the number of teams with more favourable odds in the group) are used with the aim of contextualising the performance of certain teams, in order to interpret the impact of their risk-seeking in a more intricate manner. To this end, Figure 10 plots the *rDP* and *RVE* of each team in the Group Stage, where the quadrants are interpreted in the same manner as in Figure 8, and the dividing lines are the median of the Group Stage *rDP* and *RVE*.

Certain teams are emphasised upon: AUS (+400, 3), MAR (+200, 2), and BEL (−1000, 0). These odds suggest that BEL are heavy favourites to qualify from their groups whereas MAR are not, yet Figure 7 shows that BEL plays three games indicating an elimination in the Group Stage. In Figure 10, it can be seen that the team which qualifies instead of BEL in its group, MAR, outperforms the former in terms of *RVE* (more efficient passes under pressure), while being only slightly less risk-seeking than BEL.

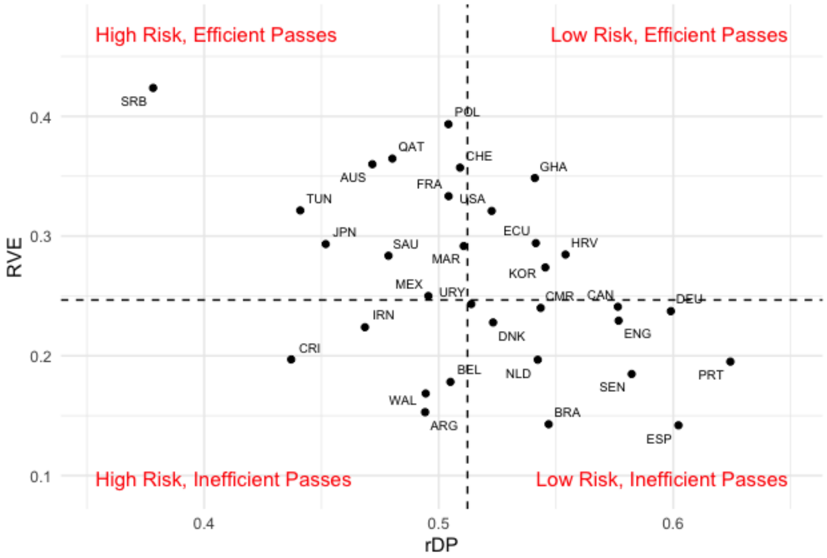


Figure 10: Scatterplot of *rDP* and *RVE* values in the Group Stage for each team

Moreover, AUS, with the worst odds of qualifying from its group, likewise outperforms its group-mates TUN (+100, 2) in *RVE*, and DNK (−333, 1) in both metrics; TUN, with worse odds than DNK, still finish higher in their group, having similarly outperformed DNK in these metrics. Moreover, AUS finish the group-stage level on points with the fourth member of the group FRA, wherein the former outperforms the latter in both metrics in the Group Stage. The above cases serve to indicate that some underdogs may find success by adopting an aggressive approach with respect to passes under pressure, especially if those passes are directed towards more valuable zones of the pitch.

To determine whether there is a statistically significant relationship between the final ranking of a team and its *RVE* value across the tournament, Figure 11 is utilised. Points which lie near or on the dashed 45-degree line represents teams with an *RVE* rank that corresponds to their actual rank. Teams that lie below this line have a lower *RVE* rank than their final position in the tournament; for instance, FRA would be expected to finish 13th in the tournament based on their *RVE*, but in reality placed 2nd. Similarly, teams lying above the line are expected to place better based on *RVE* than what is truly the case; the red regression line is not significant, indicating that *RVE* scores alone do not explain the final position of a team.

Therefore, an answer to the sub-question "*do teams that demonstrate a greater propensity for risk-seeking passes under pressure achieve a better rank in the World Cup?*" can be formulated. To some extent, based on the findings in Table 11, the risk profile, as measured by the *RVE* of a team does have an impact on their success in the tournament, however, this metric is not solely able to determine the amount of success that will be achieved.

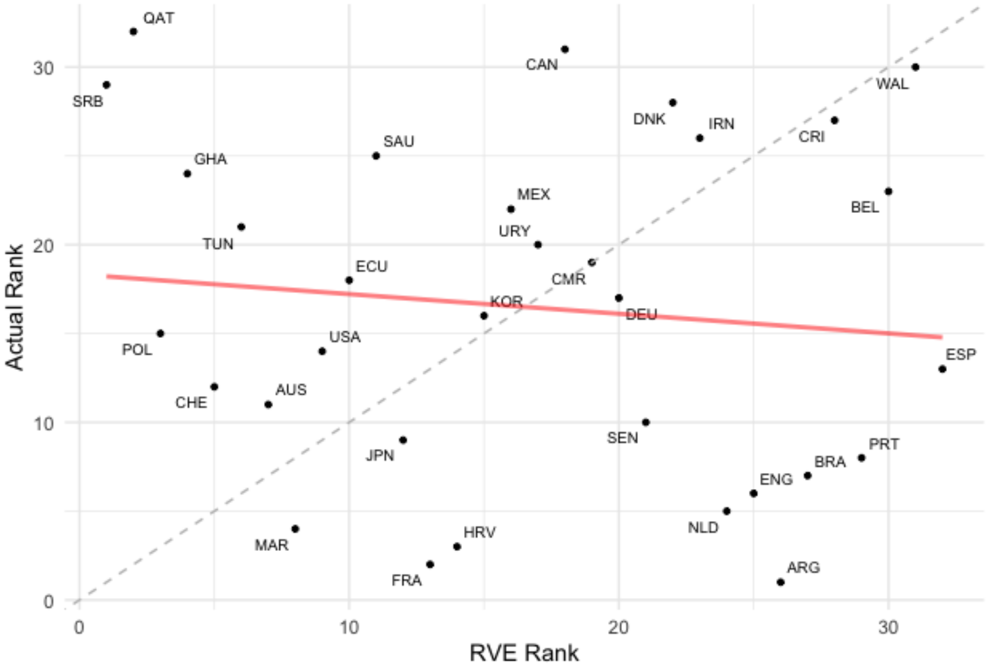


Figure 11: *RVE* rank against Actual Rank in the tournament for each team

Certain limitations of the analysis are prevalent. Firstly, as the counterfactual passes do not occur in reality, it cannot be said with certainty whether they would be intercepted by the opponent or by the team itself. Therefore, the probabilities of these passes being intercepted which are estimated by the model cannot be verified; therein lies a major limitation of this work.

Secondly, the lack of important features such as the body orientation with which players make and receive passes, or additional players being present in a passing situation but not seen in the broadcast camera are all factors that impact the extent to which the risk of a pass can be reliably estimated. Information such as body orientation and player speed, whilst possibly accessible elsewhere, are not featured in this dataset. Therefore, their inclusion is beyond the scope of this analysis.

Lastly, the limited sample size impedes the construction of approaches that can appropriately account for the playing style of a team when valuing the passes they make. As a result, measures of value such as the EPV are relied on, which, while not entirely unsuitable, are not optimal either. Being able to capture a greater diversity of playing styles provides additional context to enrich the research and the conclusions that can be drawn from it.

6 Conclusion

This research aims to answer the following research question: *to what extent do football teams demonstrate risk-aversion in their passing behaviour when under pressure?* Performing the analysis indicates that teams seem to prioritise risk-minimisation for passes under pressure at the expense of moving the ball to more valuable zones on the pitch. For teams with players of high quality, the decision to minimise risk is potentially deliberate, as these teams choose to find said players more often than their other teammates who may be in more valuable zones of the pitch. Accordingly, underdog teams - those teams that possibly lack such players - can profit by adopting aggressive strategies which prioritise taking risks and finding players in valuable zones with passes under pressure.

Moreover, a statistically significant relationship between risk-seeking behaviour and tournament ranking is not found. Teams that progress from one round to another in the tournament do not show statistically significant differences in their risk-seeking behaviour relative to teams that are eliminated in that round. Nevertheless, the analysis suggests that loss-aversion has a dominating effect; teams that progress furthest in the tournament demonstrate conservative decision-making with respect to risk-seeking for passes under pressure until competing in the final stages of the tournament. This result provides evidence for teams waiting until the final moments of a game or tournament before adopting riskier approaches; it is preferable to avoid losing with a riskier approach earlier in the tournament than later. Furthermore, for games that end in penalty shootouts, teams that show less risk-aversion during the game than their opponent tend to win in the shootout. This observation suggests that decision-making involving risk is affected by pressure, and teams which are confident in their decision-making under pressure express more risk-seeking behaviour that may translate to greater success.

Thus, when under pressure, football teams tend to choose risk-averse strategies which are consistent with prospect theory to a large extent in their passing behaviour. With respect to practical applications, there are some insights that can be drawn from this research in terms of tactics and strategies, as it has been demonstrated that teams do not necessarily need to assume an increased amount of risk when making passes under pressure to consistently find players in valuable areas of the pitch from these passes. Similarly, as discussed above, underdog teams can greatly benefit by adopting aggressive and efficient passing strategies consistently during the tournament.

In terms of theoretical contributions, this research demonstrates one of the vast possibilities of applying the novel StatsBomb (2021) data, and provides a foundation for analysing counterfactual situations across football matches. The manner in which machine learning models are used in this research supplements the existing knowledge regarding their performance on imbalanced datasets. Furthermore, a framework is created within which prospect theory can be studied across sporting tournaments, thereby expanding upon sports psychology and decision-making that occurs under pressure. That is, the methods applied in this research can be used for analysing other football tournaments and investigating whether similar results are found. The research provides a gateway to examine whether attitudes toward risk are consistent over time and tournaments, and the manner in which they develop.

Future research can consider investigating the role that the culture of a country plays in its attitudes towards risk. Additionally, research can also focus on valuing zones on a pitch based on the style of play for a team. This approach is likely to present an increasingly accurate analysis of counterfactual situations and team-specific behaviours. From a psychological perspective, newer research can look to quantify the relationship between decision-making under pressure and penalty shootout success.

References

2022 world cup odds. (2022). Retrieved from https://www.si.com/betting/2022/11/16/odds-groups-2022-world-cup

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004, jun). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, *6*(1), 20–29. Retrieved from https://doi.org/10.1145/1007730.1007735 doi: 10.1145/1007730.1007735

Bendickson, J., Solomon, S. J., & Fang, X. (2017). Prospect theory: The impact of relative distances. *Journal of Managerial Issues*, 155–168.

Bransen, L. (2017). Valuing passes in football using ball event data..

Bransen, L., Robberechts, P., Van Haaren, J., & Davis, J. (2019). Choke or shine? quantifying soccer players' abilities to perform under mental pressure. In *Proceedings of the 13th mit sloan sports analytics conference* (pp. 1–25).

Brechot, M., & Flepp, R. (2020). Dealing with randomness in match outcomes: How to rethink performance evaluation in european club football using expected goals. *Journal of Sports Economics*, *21*(4), 335–362.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, *45*(1), 12-19. Retrieved from https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199401%2945%3A1%3C12%3A%3AAID-ASI2%3E3.0.CO%3B2-L doi: https://doi.org/10.1002/(SICI)1097-4571(199401)45:1⟨12::AID-ASI2⟩3.0.CO;2-L

Burriel, B., & Buldú, J. M. (2021). The quest for the right pass: Quantifying player's decision making. In *StatsBomb Innovation in Football Conference, London, United Kingdom.*

caret: Classification and regression training. (2023). Retrieved from https://cran.r-project.org/web/packages/caret/index.html

Cefis, M. (2022). Football analytics: A bibliometric study about the last decade contributions. *Electronic Journal of Applied Statistical Analysis*, *15*(1), 232–248.

countrycode: Convert country codes. (2023). Retrieved from https://rdocumentation.org/packages/countrycode/versions/1.5.0

Endgame: Lionel messi's quest for world cup glory. (2022). Retrieved from https://theanalyst.com/eu/2022/12/lionel-messi-world-cup-stats-argentina-2022-final/

Faraway, J. (2010). Generalized linear models. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education (third edition)* (Third Edition ed., p. 178-183). Oxford: Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780080448947013312 doi: https://doi.org/10.1016/B978-0-08-044894-7.01331-2

Fernández, J., Bornn, L., & Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In *13th MIT Sloan Sports Analytics Conference.*

Fifa world cup final rankings: List of teams by record and finish at qatar 2022 from worst to first. (2022). Retrieved from https://www.sportingnews.com/us/soccer/news/fifa-world-cup-rankings-2022-final-list-teams-record-finish/yvwh7l1urgugycsmmxqflp7y

Fox, J. (2015). *Applied regression analysis and generalized linear models.* Sage Publications.

Friends of tracking data. (2021). Retrieved from https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Fritz, M., & Berger, P. D. (2015). Chapter 11 - will anybody buy? logistic regression. In M. Fritz & P. D. Berger (Eds.), *Improving the user experience through practical data analytics* (p. 271-304). Boston: Morgan Kaufmann. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780128006351000112 doi: https://doi.org/10.1016/B978-0-12-800635-1.00011-2

García-Aliaga, A., Marquina, M., Coteron, J., Rodriguez-Gonzalez, A., & Luengo-Sanchez, S. (2021). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science & Coaching*, *16*(1), 148–157.

García, V., Sánchez, J., & Mollineda, R. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13-21. Retrieved from https://www.sciencedirect.com/science/article/pii/S0950705111001286 (Special Issue on New Trends in Data Mining) doi: https://doi.org/10.1016/j.knosys.2011.06.013

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In

Handbook of the fundamentals of financial decision making: Part i (pp. 99–127). World Scientific.

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an nfl game. *Journal of Quantitative Analysis in Sports*, *10*(2), 197–205. Retrieved 2024-01-16, from `https://doi.org/10.1515/jqas-2013-0100` doi: doi:10.1515/jqas-2013-0100

McKnight, P. E., & Najab, J. (2010). Mann-whitney u test. In *The corsini encyclopedia of psychology* (p. 1-1). John Wiley & Sons, Ltd. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524` doi: https://doi.org/10.1002/9780470479216.corpsy0524

Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *Plos one*, *18*(4), e0282295.

Moore, B. B., Adams, R. D., O'Dwyer, N. J., Steel, K. A., & Cobley, S. (2017). Laterality frequency, team familiarity, and game experience affect kicking-foot identification in Australian football players. *International Journal of Sports Science & Coaching*, *12*(3), 351–358.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, *24*(1), 87–103.

Mundstock, F. B., da Silva Maia, F. H., & Bicalho, C. C. F. (2021). Goal difference relationship between the national leagues of Brazil, Germany and England from the perspective of the prospect theory. *Journal of Physical Education and Sport*, *21*(5), 2569–2575.

Peeters, T., & van Ours, J. C. (2021). Seasonal home advantage in english professional football; 1974–2018. *De Economist*, *169*(1), 107–126.

Pollard, R. (2006). Worldwide regional variations in home advantage in association football. *Journal of sports sciences*, *24*(3), 231–240.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. Retrieved from `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301` doi: https://doi.org/10.1002/widm.1301

Pulis, M., & Bajada, J. (2022). Reinforcement learning for football player decision making analysis. In *Statsbomb conference.*

Riedl, D., Heuer, A., & Strauss, B. (2015). Why the three-point rule failed to sufficiently reduce the number of draws in soccer: An application of prospect theory. *Journal of Sport and Exercise Psychology*, *37*(3), 316–326.

Rufibach, K. (2010). Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, *63*(8), 938–939.

Ruiz, A., & Villa, N. (2008). Storms prediction: Logistic regression vs random forest for unbalanced data. *arXiv preprint arXiv:0804.0650*.

Saito, T., & Rehmsmeier, M. (2015, 03). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), 1-21. Retrieved from `https://doi.org/10.1371/journal.pone.0118432` doi: 10.1371/journal.pone.0118432

Schauberger, G., & Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, *18*(5-6), 460-482. Retrieved from `https://doi.org/10.1177/1471082X18799934` doi: 10.1177/1471082X18799934

Sha, L., Lucey, P., Zheng, S., Kim, T., Yue, Y., & Sridharan, S. (2017). Fine-grained retrieval of sports plays using tree-based alignment of trajectories. *arXiv preprint arXiv:1710.02255*.

soccermatics package. (2019). Retrieved from `https://rdocumentation.org/packages/soccermatics/versions/0.9.5`

StatsBomb 360 Data. (2021). Retrieved from `https://statsbomb.com/news/statsbomb-announce-the-release-of-free-statsbomb-360-data-euro-2020-available-now/`

StatsBomb Data Specification v1.1. (2019). Retrieved from `https://github.com/statsbomb/open-data/blob/master/doc/StatsBomb%20Open%20Data%20Specification%20v1.1.pdf`

Van Roy, M., Robberechts, P., Yang, W.-C., De Raedt, L., & Davis, J. (2021). Leaving goals on the pitch: Evaluating decision making in soccer. *arXiv preprint arXiv:2104.03252*.

Wikipedia contributors. (2024). *List of players who have appeared in the most fifa world cups — Wikipedia, the free encyclopedia.* `https://en.wikipedia.org/w/index.php?title=List_of_players_who_have_appeared_in_the_most_FIFA_World_Cups&oldid=1200371348`. ([Online; accessed 22-February-2024])

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, *77*(1), 1–17. Retrieved from `https://www.jstatsoft.org/index.php/jss/article/view/v077i01` doi: 10.18637/jss.v077.i01

A ISO3C Codes for Country Names and Description of Code Files

Table 12: Original Country Names and ISO3C Codes.

Country Name	ISO3c Code
Argentina	ARG
Australia	AUS
Belgium	BEL
Brazil	BRA
Cameroon	CMR
Canada	CAN
Costa Rica	CRC
Croatia	HRV
Denmark	DNK
Ecuador	ECU
England	ENG
France	FRA
Germany	DEU
Ghana	GHA
Iran	IRN
Japan	JPN
Mexico	MEX
Morocco	MAR
Netherlands	NLD
Poland	POL
Portugal	PRT
Qatar	QAT
Saudi Arabia	SAU
Senegal	SEN
Serbia	SRB
South Korea	KOR
Spain	ESP
Switzerland	CHE
Tunisia	TUN
United States	USA
Uruguay	URY
Wales	WAL

A brief explanation of the attached code files:

1. Pass Data: file that loads all the events of the tournament.
2. int_prediction_cv: Models that predict probabilities of each opponent in the frame intercepting passes. Contains code for calculating the probability of each pass being intercepted by opponents with selected model (Random Forest).
3. possession_prediction_cv_updated: Models that predict probabilities of teammates in the frame of the pass intercepting the ball. Contains code for calculating the probability of each pass being intercepted by the actor's team with selected model (Random Forest), and to obtain the risk of a pass.
4. duration_prediction: Random Forest model that is trained to predict the duration of a pass.
5. counterfactual_test: Calculate the counterfactual pass variables, predict the probability of a counterfactual pass being intercepted by every opponent and teammate in the frame using selected models, and estimate risk of a counterfactual pass.
6. ZPV: Calculate the ZPV of original and counterfactual passes.
7. quantifying: Obtain the rDP, vDP, and RVE for each team during the tournament.
8. rdp_plots and rve: Create plots for rDP and RVE results and other miscellaneous analysis.