ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis: Data science and Marketing Analytics

Analyzing the impact of substitutions in football matches

Quint van Leeuwen

Student ID: 530790

Supervisor: Dr. Phyllis Wan

Second assessor: Dr. Vardan Avagyan

August 2020

Abstract

This Master's Thesis analyzes the impact of substitutions within the game of football. Event data from seven top leagues is used from the 2017/2018 season. The most important predictor variables for the classification task of all match outcomes were found. The four different performed classification methods were: Ordinal Logistic Regression, Random Forest, Support Vector Machines and Naive Bayes. As a result, the Random Forest method appeared to be the best predicting and interpretable method. Offensively, the resulting most important variables were the 'Expected goals' and the 'Shots on target' metrics. Defensively, the most important variables were the 'Duels won' and 'PPDA' metrics. Finally, by a panel data analysis, the substitution impact on respective variables was investigated. Depending on the scoreline scenario, on average, a statistically significant impact is made on the variables by substitutions. On average, top-ranked teams' substitutions seem to have a greater impact than low-ranked teams' substitutions. Within the final section of this thesis, the late-scoring teams' substitution strategy was also investigated. It was observed that late-scoring teams, on average, utilize their offensive substitutions earlier and substitute more offensively than its opponents.

Acknowledgements

Foremost, I would like to express my gratitude to my supervisor dr. Phyllis Wan for supervising me during my Master's Thesis process. By her thorough methodological knowledge, she provided me with detailed feedback and advice. Finally, I would also like to thank my family for all their support during these past months.

Key Words

Substitutions, Impact, Football, Coaches, Ordinal Logistic Regression, Random Forest, Support Vector Machines, Naive Bayes, Panel data analysis, Linear Regression, Poisson Regression.

Contents

List of Tables

List of Figures

1. Introduction

This section will provide the main motivation and goals for this research. Furthermore, the research questions will be formulated and the contributions to the field of football analytics will be explained along with its relevance and application. Finally, this thesis' structure will be described.

1.1 Research context

All football fans probably recall watching a match when a substitute immediately made a big impact on the game, perhaps by scoring a critical goal or blocking a dangerous offensive shot. Was this managerial brilliance in terms of knowing when to substitute? Perhaps it is simply a case of memory bias and confirmation bias **(Schacter, 1999)**. In sports, people tend to remember outstanding events and use these occasions to solidify previously held opinions **(Silva et al., 2016)**. Also, the opposite could occur, where a substitution strategy decreases team performance. Since the resumption of the German Bundesliga after the intermission caused by the COVID-19 virus outbreak, it was allowed to use five substitutions per match instead of three. This led to the first time in modern football history where a coach made four substitutions at once. It was FC St. Pauli's coach Jos Luhukay whose team was trailing a 1-0 score. In the 61^{st} minute he did his trick and substituted four fresh players at once to try to turn the tide. Unfortunately for FC St. Pauli, the substitutes did not have a positive impact on the game and after the moment of substituting, their opponent Darmstadt 98 scored three goals in the remaining 30 minutes of the game. It led to a big loss for FC St. Pauli, with a final score of 4-0.

This thesis will aim to shed light on the *quantitative impact* of substitutes by analyzing historic match data. One could wonder if substitution strategies exist that systematically outperform others, or if substitutions made by top-ranked teams have a greater in-game impact. Thus, this thesis will aim to investigate the effectiveness of substitutes in football.

1.2 Problem description

By substituting, for other reasons than injuries or killing time, the coach wishes to positively influence the evolving match situation. Potentially, the coach wishes to change the current team strategy into more offensive or defensive, or he wishes to replace players because of build-up fatigue during the match. Hence, the primary function of a coach during a match is to adapt his team to the evolving situation by deploying and managing the limited resources at his disposal either in

order to gain or maintain an advantage or to retrieve a losing situation. The essential question remains whether substitutions can really lead to a competitive edge if done successfully and whether that can be backed up statistically. When comparing football to other sports such as basketball, it is generally more difficult for the coaches to have an influence on the match outcome once the match has started. One reason is the pace and flow of the game of football. But another reason is that there are almost no stoppages for the coach to discuss the team strategy with its players and to make flashy draws on clipboards. Therefore, the most critical in-game decision that a coach has in football is the utilization of his three substitutions. In the world of football, disagreement exists about the substitutions' impact or about which substitution strategy works best. Hence, are coaches actually able to make a significant match impact by utilizing their substitutions? And do some substitution strategies exist that structurally outperforms others?

Historic match data can be used to answer such questions. Lately, the integration between the game of football and the use of data has been on a rise, where more clubs are shifting towards a data-driven approach (**Bornn et al., 2018**). While small clubs cannot compete on budget, they can make smart, data-informed decisions to help close the gap on their heavy-spending competitors, levelling the playing field – as demonstrated by the Oakland Athletics baseball team, whose success through analytics was documented in the movie *Moneyball* (**Bornn et al., 2018**). Since the digital transformation is affecting the world of sport the developments in the field of sensor technology have led to a rapid increase in the volume of data. In particular, spatio-temporal position data, which are now available almost everywhere in football, harbour huge potential for performance analysis (**Link, 2018**). The challenge is to turn this data into advantageous insights that augment and enhance the knowledge about the game of football. Surely, once the data is analyzed effectively by the club's general management or coach, it can work to a club's advantage on the playing field. A competitive edge can be gained by utilizing and interpreting the data correctly. Especially low-budget teams could reap benefits from such data-driven findings since their budgets usually do not allow them to buy the best quality players.

1.3 Research questions

Firstly, this thesis addresses several general research questions, whereafter also more specific research questions will be answered. First of all, it is important to assess the key performance indicators that help explain a team's performance. Therefore, the first research question is:

1) Which variables are most important when measuring team performance of football teams?

The second phase of this research will be to investigate the impact of different substitution strategies on the previously found team performance indicators. We can distinguish substitute strategies into the types: offensive, neutral, and defensive. The impact of offensive substitutions will only be measured on the important metrics which describe offensive aspects of the game, while the impact of defensive substitutions will be measured on the important metrics which describe defensive aspects of the game. This leads to the second and third research questions:

2) Do offensive substitute strategies have a significant impact on the previously found important variables which describe the offensive aspects of the game?

3) Do defensive substitute strategies have a significant impact on the previously found important variables which describe the defensive aspects of the game?

Subsequently, it will also be researched on which variable the substitutes' impact is greatest. Thus, the fourth research question is:

4) If any, on which variable(s) is the substitutes' impact the greatest?

After previous general research questions, a more specific approach will be handled. Due to bigger budgets, higher prestige, and national and international successes, top-ranked European teams are considered to have a much better and also wider player selection. Because a coach is limited to lining up 11 players every match, a group of top-level players, who can really make a difference, are positioned on the bench as a substitute. This thesis will investigate the substitutes' impact of top-ranked European (league) teams and the substitutes' impact of low-ranked teams and the potential difference will be compared. Thus, the fifth research question is:

5) Do top-ranked teams, in general, have a greater substitution impact?

Finally, some teams are frequent scorers of late goals, where 'late' is defined as approximately after the 70^{th} minute. This thesis aims to analyze the substitution strategies of respective teams to find out if these teams' substitution strategy is related to frequently scoring late goals. Subsequently, this gives insights into whether respective coaches can be regarded as brilliant strategists or not. Thus, the final research question can be formulated as:

6) Do late-scoring teams, on average, hold onto a different substitution strategy?

Hereby, both offensive and defensive strategies will be researched along with the timing of the respective strategy adjustments.

1.4 Contributions

Previously performed research on the impact of substitutions mainly focusses on either the individual players' work-rate or on the substitution impact on the match outcome (earned points). In contrast to other research, this research will evaluate the substitutes' impact measured on several key performance indicators, such as 'Expected goals' (defined in Section 4.2). Meanwhile, this research tries to shed light on the difference in substitution impact between top-ranked and low-ranked teams. And finally, the substitution strategies of late-scoring teams are explored. Thereby, within the field of football analytics, new areas are being explored and researched.

1.5 Relevance and application

This thesis aims to bring more clarity to the team performance impact that coaches can make by using their substitutions and provides general insights about key performance indicators in football as well. Data-driven insights, indicating the impact of substitutes on the match outcome under different scoreline scenarios, are useful for either the coach or the general management of the club. Namely, questions can be answered whether it would be effective to substitute offensively at a given scoreline? By using statistical back-up, a coach is able to maximize his team's tactical advantage, without solely basing his substitution strategy on "gut feeling". Hence, this thesis sheds light on the type of substitutions, implying the effectiveness of either offensive or defensive substitutions. Besides that, research is conducted about the difference between substitutions of top-ranked versus low-ranked teams. Thereby, the general management of clubs could derive meaningful insights about the importance of a wide player selection, which top-ranked teams typically have. The latter research topic reveals the marginal gains a club can obtain by allocating much of its budget into attracting additional top-class players, who in turn, will be regularly positioned on the bench as a substitute. Perhaps, effectively spoken, the club's budget can better be allocated elsewhere. In the final stage of this thesis, the substitution strategies of 'late-scoring' teams are delved into. Thereby, it can be figured out whether there is a relation between unique substitution strategies and these teams' ability to score late in the game on a regular base.

The gap between a football or sports context to a more general business context can be bridged. The parallel can be drawn to either human resource management practices or more specifically to the medical sector for example. Namely, the productivity of a hospital's personnel can be reviewed, who continuously take over working shifts to be able to provide aid to patients at any time of the day. Like substitutes in a football context, the impact of each working shift in a hospital can be researched. Especially during the times of the current COVID-19 pandemic, there has been a big increase in the amount of stress put on the hospital staff. **Wu et al., (2009)** researched the psychological impact of the SARS epidemic on hospital employees in China and found that 10% of the employees had experienced high levels of posttraumatic stress (PTS) symptoms since the SARS virus outbreak. This research shows high amounts of stress on the hospital personnel can have a severe impact, which in turn might have its effect on their productivity. Furthermore, **Emanuel et al., (2020)** point out the scarcity of the current U.S. medical resources, and provide recommendations to optimize the allocation of healthcare during the COVID-19 outbreak. Research which, like the impact of substitutions in football, investigates the productivity of hospital personnel shifts can be performed and could provide helpful insights into finding the optimal length of the working shifts, or which type of personnel is the most effective for some specific duty. Thereby, a parallel of this research could complement **Emanuel et al.'s (2020)** recommendations. Probably, an optimization of the hospital's working shifts could reduce the amount of stress on the personnel during the COVID-19 outbreak, and thereby increase the personnel's productivity.

Regarding human resource management practices in a business context, this research could also be paralleled. Say, a company wishes to evaluate the impact of the appointment of newly-hired employees. To hire these employees, the company has had an increased cost in total salary, and the question remains whether that is paying off. Therefore, the company could wish to measure the impact of the employees on several key performance indicator variables, which is easily comparable to this research, where the marginal impact of substitutions (new employees) in football are investigated. In cases when the new employees' contribution is below expectation, the human resource department may be willing to reconsider the respective employees' appointment.

Finally, as was mentioned in the first paragraph of Section 1.1, this thesis will also question the psychological suspicion of **(Silva et al., 2016)**, who mention that substitutions' impact may be prone to memory bias and confirmation bias **(Schacter, 1999)**. Their claim that people tend to remember outstanding events and then use these occasions to solidify previously held opinions will

be evaluated in a football substitution context.

1.6 Thesis structure

Hence, to answer previously mentioned research questions, data of more than 1,900 matches is used from the top national leagues in Europe - the English Premier League, the Italian Serie A, the Spanish La Liga, the German Bundesliga, the French Ligue 1 -, and the European and World Championships of 2016 and 2018 respectively. Continuously, the set-up of this thesis is divided into two global steps.

The first step is to find the most important offensive and defensive team performance variables in terms of predicting the match outcome. The match outcome is categorical and divided into three categories: 'Lose', 'Draw', or 'Win'. To tackle this problem, several multi-class classification models will be performed and compared based on their predictive accuracy. Hereby, the goal is to obtain the variable importance and thereby observe the most important team performance indicators. Multiple classification models are suitable to tackle such problems and all models' performance can be tested on a separate test data set, which is explained in more detail later. The classification models which are used in this thesis are Ordinal Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes.

The second step will be to assess the substitutes' impact on these most important variables which have been found by step one. Offensive and defensive substitution strategies will be analyzed under different scoreline scenarios. At first, per match, a differential will be computed describing which team substituted *more offensive* or *more defensive* than its opponent, which allows us to investigate the respective team's metrics in the given match. Note that, when both teams use exactly the same substitute strategy, no conclusions about one team's substitute impact can be drawn because the potential effects cancel each other out. Therefore, such differential will be essential within this research. When the relevant subset of investigatable matches has been identified, the real-time match statistics will be computed for those matches. The data will now be time series data, with t minute intervals between subsequent observations. Finally, this allows performing a linear regression model or Poisson regression model (depending on if the explanatory variable is discrete or continuous), on the real-time match statistics, where a dummy variable is created describing whether a team has offensively or defensively substituted at a respective time interval. The coefficient and statistical significance of the respective dummy variable will be of interest to answer the research questions of

this thesis.

By doing so, based on historical match data, this research' goal is to bring more clarity to the *quantitative impact* on team performance that coaches can make by using their substitutions.

2 Literature review

After carefully studying literature on the impact of substitutions in football, some interesting findings have been identified. In this paragraph, the key takeaways from the initial research that is in line with this thesis' topic will be discussed.

2.1 Impact of substitutes

Gomez et al. (2017) claim that the substitution time strategies of a coach can have an impact on the final outcome in a match and on the playing tactics if done effectively. Also, **Myers (2012)** agrees upon that. **Hills et al. (2018)** state that the impact of substitutions is not so apparent. Evidence does exist that suggests that particularly the substitution of midfielders leads to an increase in the work-rate **(Bradley et al., 2013)**. However, the authors claim that the overall contribution of substitutes to team success remains to be determined. Because of contradicting research results, further investigation about the impact of substitutes will be relevant.

Trainor (2014) investigated the general phenomenon of fatigue amongst football players. In the world of football analytics, it has already been established that the nearer the ending of the match, the greater the goal expectation is. In Section 3.2, Figure 4 is a data visualization of the scored goals' distribution. One of the reasons why substitutes score at a higher rate per minute played than starting players is the increased goal expectation near the end of the match, typically the time when substitutes enter the pitch. **Trainor (2014)** believes that players should be substituted early in the game. He claims that coaches can make an impact on the game by effective and early substitution strategies. Beyond the fact that managers could hold a substitute back in case of injuries in the game, he sees no reason why the managers should not take the full benefit of the fresh players on the bench. His research is based on the analysis of the five top European leagues (England, Spain, Italy, Germany, France) over the 2012/2013 season, where he compared goal-scoring rates between three groups of players for each match: players who played the full 90 minutes, players who are subbed on, and players who are subbed off. On average in all leagues, the goal-scoring rate (per 90

minutes) was highest for the substitutes, followed by the players who are subbed off. The main finding was the subbed off player's goal-scoring rate is higher on average than the one of the players who played a full match. Thereby, his final conclusion is that the fatigue factor of the players looks so strong that it even overcomes the fact that most goals are scored towards the end of the match.

Mohr et al. (2003) claim that substitute players cover 25% more ground while running at high-intensity during the final 15 minutes of the match than other players. However, it must be noted that the research sample was extremely small. Another study about substitutions' work-rate in the English Premier League, found that substitute players were able to cover between 10% to 27% more distance in high-intensity running than their peers who were replaced or who played the whole game **Bradley et al. (2014)**. According to these studies the work-rate of the substitutes is significantly higher compared to other players, suggesting that substitutes are indeed able to make significant impact on the outcome of the football matches when coming onto the pitch. Therefore, the strategic use of substitutes may be able to reduce the overall team's fatigue.

Bartling et al (2015) performed a studies about expectations being reference points and applied it to a football context. The authors found that coaches implement offensive strategy adjustments, through substitutions, significantly more often if their teams are behind expectations. Besides that, **Bartling et al (2015)** claim that substituting players offensively while being behind expectations *worsens* the expected ultimate match outcome. This leads to the research's conclusion that coaches might feel pressure or frustration when being behind expectations, which can manifest itself in different and potentially not entirely rational behaviors.

2.2 Performance differences amongst player types

Midfielder substitutions are the most frequent ones (see Table 3 in Section 3.2). **Cabo et al. (2018)** researched the work-rate of football players in the Spanish La Liga in season 2014/2015. One of their key findings is that substitutes covered greater distances at high and medium intensity running intervals compared to other players who played the full match. Particularly central midfielders, and to a lesser degree forwards, increased their covered distance at high-intensity running intervals compared to players on their position who played the full match. Regarding the substitution of central defenders, fullbacks and wide midfielders such results were not found, and no differences were spotted. The high-intensity running variable is considered an important indicator of physical performance in professional football **(Cabo et al., 2018)**. The authors state that research has

demonstrated that the number of performed sprints, high-intensity running (HIR) and distance covered are lower in the second half than in the first half of the game. Additionally, **Mohr et al. (2003)** found that top-class players generally perform more HIR periods during a game than players at a less elite standard. The latter finding might be relevant concerning this thesis' fourth research question.

2.3 Timing of substitutions

Previous research about the optimal substitute times has been performed and some interesting findings will be discussed within this section. **Myers (2012)** proposed a substitution scheme based on regression tree methodology that analyzed data from the top four soccer leagues in the world: the 2009/2010 seasons of the English Premier League, the German Bundesliga, the Spanish La Liga and the Italian Serie A. Also, the 2010 season of North America's Major League Soccer (MLS) and the 2010 FIFA World Cup were analyzed. The general decision rule advocated by **Myers (2012)** about the optimal substitution times is as follows. When a team is losing, utilize the 1^{st}, 2^{nd} and 3^{rd} substitute before the 58^{th}, 73^{rd} minute and 79^{th} minute, respectively. And when a team is at tie or winning, substitute at will.

In his research, **Myers (2012)** demonstrated that teams that followed his decision rule improved their goal differential 42% of the time. For teams that did not follow the decision rule, they improved their goal differential only 21% of the time. **Myers (2012)** mentions that coaches underestimate the significance of fatigue late in a match, which causes them to overvalue starters and undervalue substitutes. Thereby, he emphasizes the importance of utilizing the substitutes early in the game, by which the players' fatigue can be adequately countered. In line with this finding, **Rey et al. (2015)** also suggest that coaches should be aware that reverting losing scenarios requires to change tactics early in the match.

Silva et al. (2016) argue that there are no special substitution times or periods of a match that yield a competitive advantage for trailing teams. Thhe authors state that there is no discernible time during the second half where a substitute leads to a clear benefit. Remarkably enough, differences amongst researchers' results are found, as **Myers (2012)**'s research claims otherwise. However, in both studies different response variables were used, so their results cannot be directly compared since the research had different approaches and data. Furthermore, a sidenote is made by **Silva et al. (2016)**, in which the authors suggest that managers should substitute, especially when a

player's performance drops (e.g. because of fatigue). They state however, that there is no need to tie the substitutions to the crucial times such as the 58^{th}, 73^{rd} and 79^{th} minute as was advocated by **Myers (2012)**.

Additionally, as was mentioned in previous section, **Gomez et al. (2017)** found that when a team is losing the coach substitutes earlier than when winning or drawing. Intuitively, this finding makes much sense since it is likely that the coach of a losing team will try to turn the tide by adjusting the team's strategy. Moreover, **Del Corral et al. (2008)** found that defensive substitutions are generally made later in the match than offensive substitutions.

3. Data description

In this section the data sets that were used for this research will be provided. To provide initial insights into the data, the data sets will be explored along with visualizations and summary statistics. Additionally, it is explained how the data is split into a training set, a validation set and test set. Finally, some limitations of the available data sets are discussed.

3.1 Data sets

The data sets that are used within this research are collected by the company Wyscout. The total information comprises 1,941 football matches, 3,251,294 events and 4,299 players. The data collection is done by expert video analysts that use a tagger device combined with the appropriate software. For the tagging of all in-game events, three operators are involved to guarantee the accuracy of the data collection. Each team has one operator, and the third operator is supervising the output of the whole match. Then for each in-game event, the operator creates a new event on the timeline and adds a type, subtype, the coordinates on the pitch and all additional attributes to the event. From all the information gathered by Wyscout, the following data sets are considered for this research:

`Matches Data`

This data set comprises match data of the 2017/2018 season of the five top national leagues in Europe - the English Premier League, the Italian Serie A, the Spanish La Liga, the German Bundesliga, the French Ligue 1 -, and the European and World Championships of 2016 and 2018, respectively. Thus, in total, this research is based on match data of seven different leagues. The data set contains

information about all matches played within mentioned competitions. Of the 1,941 played matches all results are stored in combination with the line ups of both teams, and all substitutions (describing the players involved and the minute of the substitution). Table 1 provides a brief overview of the available variables.

Table 1: Variable description Matches Data

Variables	Description
Match ID	Unique identifier for each match
Team ID	Unique identifier for each team
Player ID Substitutes	Unique identifier of substituted players
Label	Contains the name of the two clubs and the result of the match
Home/Away Team	Categorical variable describing which team played at home and away
Score	Numerical variable describing the scored goals per team
Time of Substitutes	Numerical variable describing the minutes each substitute took place

`Events Data`

This data set contains information about all the events that occurred during each match. Seven types of events exist that were registered during these matches. The most important events types are presented in Table 2, and comprise the main data of `events data` that occur in the data set along with previously mentioned variables like 'Match ID' and 'Team ID'. Furthermore, these event types are accompanied by the corresponding time of occurrence, players involved, position on the field and a tag describing additional information about the event (e.g. if the pass was accurate).

Table 2: Variable description of most important events types

Type	Subtype	Tags
Pass	Cross, simple pass, launch, smart pass, head pass	Accurate, not accurate, key pass, opportunity, assist, goal
Foul		No card, yellow, red, second yellow
Shot		Accurate, not accurate, block, opportunity, assist, goal, position: goal, position: out, position: post
Duel	Air duel, defending duel, attacking duel, ground loose ball duel	Accurate, not accurate, won/lost, sliding tackle
Free kick	Corner, free kick shot, goal kick, throw in, penalty, simple kick	Accurate, not accurate, key pass, opportunity, assist, goal
Others on the ball	Clearance, touch	Interception

The positions of the events are given in normalized (x, y)-values as coordinates, where x reflects the relative distance to the opponent's goal on a $[0, 100]$ scale, while y is the event's nearness to the right side of the pitch on a $[0, 100]$ scale.

`Players Data`

Additionally, the data set `players data` was used to extract valuable information about the players who were involved in the substitutions. This data set comprises information about the players playing in all seven leagues which were previously metioned. The information that will be used within this research is the team each player plays for, the national team each player plays for (if the player is selected for its national team) and the player's main field position (defender, midfielder or forward).

3.2 Data exploration

Within this section the basic statistics about the data set are provided. The data set `events data` consists of a set of observations that each have a different type, depending on the sort of event it describes within a match. Figure 1 shows the percentual occurrance of the types of all observations of all seven leagues.

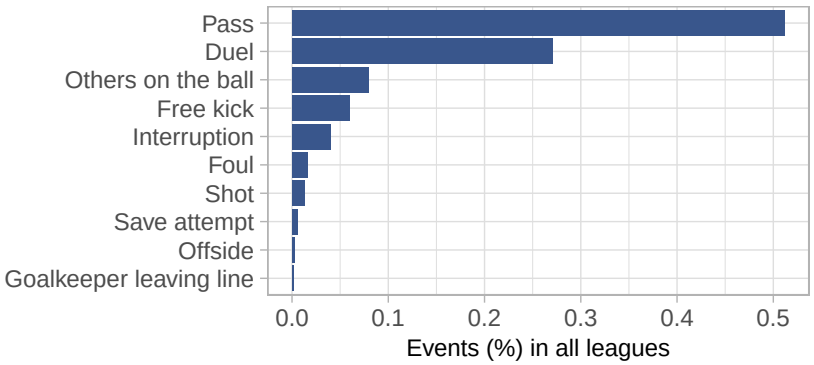


Figure 1: Events distribution

As can be observed in Figure 1, the majority of the events consists of passes ($> 50\%$), and duels ($> 25\%$), whilst only a small percentage of events consists of shots ($< 2\%$).

Beyond the type of events, the frequency of the total number of events per match are presented in Figure 2.

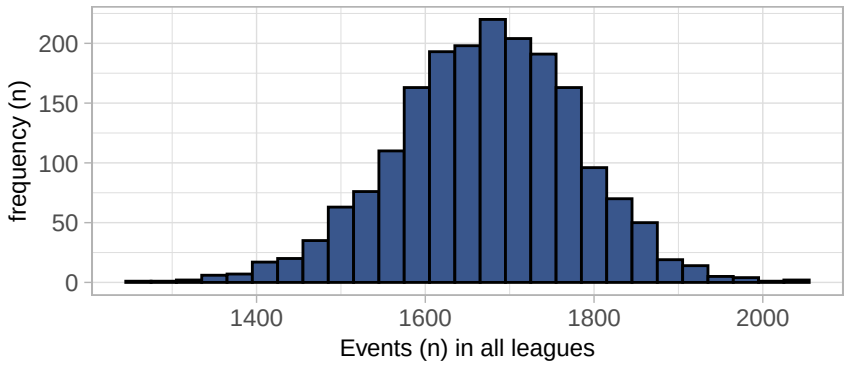


Figure 2: Distribution of the number of events in football matches

From Figure 2, it can be observed that for a great share of matches, `events data` contains between 1650 and 1750 observations of individual match events.

In Figure 3, three histograms are presented which show the timing of all made substitutes in the data set. It can be observed that coaches often utilize their first substitute at half-time (45^{th} playing minute).

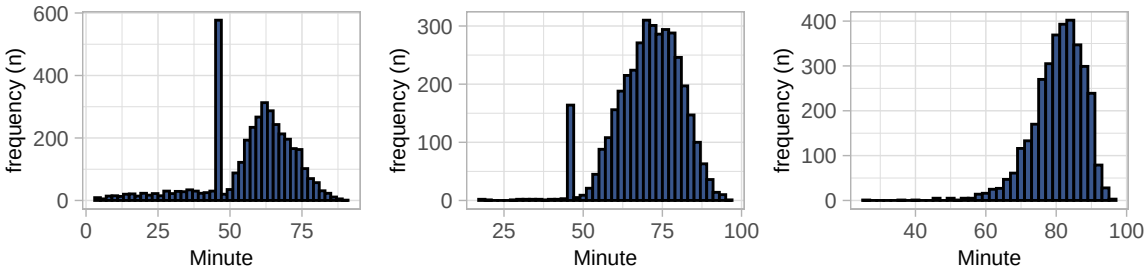


Figure 3: Histograms of the timing of first, second and third substitutions

Table 3 presents the counts of forwards, midfielders, defensive, goalkeepers substitutes within the data set on hand.

Table 3: Types of substitutes

Substitution	Defenders	Midfielders	Forwards	Goalkeepers
In	1767	5010	4202	42
Out	1710	5540	3738	33

In Figure 4, the frequency of all scored goals per playing minute are illustrated.

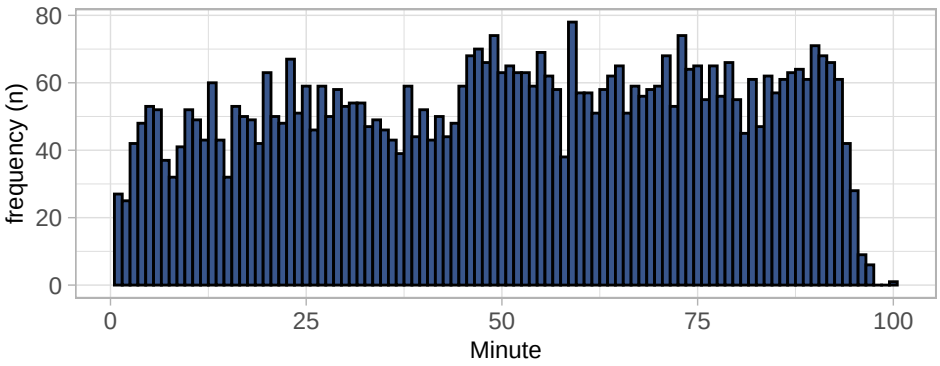


Figure 4: Histogram of minutes of all scored goals

To answer the fourth research question, teams will have to be distinguished between top-ranked and low-ranked. Therefore, the three teams that ended in the top 3 during the 2017/2018 season are considered as top-ranked. Whereas, the three teams that ended in the last 3 league positions during the 2017/2018 season are considered as low-ranked. As previously mentioned, two separate obtained data sets comprise match data from either the World Championships or the European Championships. Due to the nature of the research question, these league's match data sets will not be used. This is because, commonly, all national teams have a very wide player selection (weakening the assumption that low-ranked teams do to a lesser degree), and besides that, it is also difficult to rank the national teams in hierarchical order. In Table 4, for each national league, the top-ranked and low-ranked teams are listed.

Table 4: Top-ranked and low-ranked teams per national league

Position	England	Italy	Spain	Germany	France
First	Manchester City	Juventus	FC Barcelona	Bayern München	PSG
Second	Manchester United	Napoli	Atletico Madrid	FC Schalke 04	AS Monaco
Third	Tottenham Hotspur	AS Roma	Real Madrid	Hoffenheim	Olympique Lyonnais
...	...	...	...	...	...
Second before last	Swansea City	FC Crotone	Deportivo la Coruña	Wolfsburg	Toulouse FC
First before last	Stoke City	Hellas Verona	UD Las Palmas	HSV	Troyes AC
Last	West Bromwich Albion	Benevento Calcio	Malaga CF	FC Köln	FC Metz

Finally, related to the last research question about analyzing the substitution strategy of late-scoring teams, within this research a team will be defined as 'late-scoring' if the team frequently scores goals after the 67^{th} minute. Namely, the average *offensive* substitution moment of all teams is during the 67^{th} minute. In Section 3.4 the definition of an offensive substitution is explained in detail. Table 5 shows which teams are eligible to investigate, since the ratio of scored goals after the 67^{th} minute compared to the total number of scored goals is presented. The average *defensive* substitution moment of all teams is during the 70^th minute. However, the 67^{th} minute will be chosen as the boundary within this analysis, because offensive substitutes are more likely to be related to late-scored goals. To fully answer the research question on hand, also the defensive substitution moments of the late-scoring teams will be analyzed and compared with the defensive substitution moments of the other teams. Table 5 presents the 25 latest-scoring teams, and in Section 6.4 the research question on hand will be answered.

Table 5: The 25 latest-scoring teams

Team	Goals.after.67min	Total.goals	Ratio	avg.off.sub	avg.def.sub	avg.team.offsub.dif	avg.sub.dif
Russia	5	10	0.50	62	71	-0.25	-0.17
SM Caen	13	27	0.48	64	67	-0.07	-0.45
Belgium	12	25	0.48	71	64	0.08	-0.29
AFC Bournemouth	21	45	0.47	67	67	0.42	0.09
Crystal Palace	20	45	0.44	65	75	-0.01	-0.37
Levante	19	43	0.44	61	58	0.43	0.38
CD Leganés	15	34	0.44	66	74	0.24	0.03
Real Betis	26	59	0.44	66	74	0.67	0.70
Croatia	6	14	0.43	86	80	-0.29	-1.29
Genoa CFC	14	33	0.42	59	69	0.55	0.55
Everton	18	43	0.42	58	70	0.05	-0.15
Deportivo Alavés	16	40	0.40	69	74	0.16	0.09
FC Köln	14	35	0.40	64	64	0.12	0.09
FC Augsburg	17	43	0.40	69	68	0.16	0.07
PSG	42	107	0.39	75	68	-0.11	-0.34
Lille OSC	16	41	0.39	57	61	-0.20	-0.47
Burnley FC	14	36	0.39	71	72	0.95	0.64
SD Eibar	17	44	0.39	69	74	0.08	-0.24
Hannover 96	17	44	0.39	62	77	0.66	0.75
Watford	17	44	0.39	62	68	0.39	0.25
Rennes	19	50	0.38	64	70	0.20	-0.01
West Ham United	18	48	0.38	62	69	0.39	0.21
UD Las Palmas	9	24	0.38	63	64	0.64	0.50
UC Sampdoria	21	56	0.38	57	64	0.25	0.17
Atalanta Bergamo	21	57	0.37	65	72	0.04	0.04

Note:
Only teams which scored at least 10 goals in total are considered. Furthermore, the column 'avg.off.sub' represents each team's average moment (playing minute) of offensively substituting, the column 'avg.def.sub' represents each team's average moment of defensively substituting. The column 'avg.team.offsub.dif' represents each team's individual average substitution differential (defined in Section 3.4). Whereas, column 'avg.sub.dif' represents the matches' average substitution differential, which also takes into account the substitution differential of the opposing team, and indicates whether a team substituted more offensively/defensively than its opponent.

To get a better general insight in the data sets on hand, summary statistics of all variables over all matches per team are given in Table 6.

Table 6: Summary statistics over all matches per team (N = 3,862)

Variables	Mean	Std.dev.	Min	Max
Goals	1.35	1,26	0	8
Shots	11.65	4.95	0	39
Shots on target	4.37	2.56	0	15
Corners	4.97	2.80	0	19
Crosses	16.03	7.56	0	62
Passes	428.27	127.06	121	1008
Pass accuracy (%)	0.82	0.05	0.58	0.94
Ball possession (%)	0.50	0.12	0.16	0.84
Yellow cards	1.96	1.41	0	9
Red cards	0.09	0.29	0	2
Fouls	13.13	4.33	1	31
Duels won	87.25	16.17	41	159
Free kicks	14.54	4.55	2	33
Penalties	0.15	0.38	0	3
Invasion index	0.04	0.01	0.007	0.11
Acceleration index	0.03	0.05	0.0001	0.72
Expected goals	1.31	0.79	0	6.47
PPDA	6.65	4.73	0.85	57.92
Passing indicator	15.69	5.61	3.19	44.42
Substitutions	2.84	0.42	0	3
Offensive substitutions	0.68	0.74	0	3
Defensive substitutions	0.58	0.69	0	3
Substitution differential	0.21	0.95	-3	3

Note:
For a total of 1,931 matches the data is composed for both the home and away teams, which leads to N = 3,862

3.3 Composition of match time intervals

When performing the second step of the analysis, the substitutes' impact on the important variables will be assessed. Per match, the team's match statistics data will be evaluated real-time (time series), where for each time interval an observation is computed. Each match is divided into t time intervals. Within this research $t = 30$, meaning that each time interval will be approximately equal to $\frac{90}{30} = 3$ minutes, slightly variating per match depending on how much extra time was given by the referee. Thus, for each match 30 observations are collected, representing the real-time team performance on the respective variables per every 3 minutes played.

3.4 Substitution quantification

To be able to distinguish between offensive and defensive substitution strategies, a *substitution differential* will be computed for each team per match. The differential represents the offensiveness of a team's total substitutes. At the start of a match, when there are no substitutions made yet, the differential will equal a value of zero. During the match, the differential will be computed as follows:

- Whenever a forward is substituted instead of a non-forward, the measure will be increased by + 1
- Whenever a midfielder is substituted instead of a defender, the measure will be increased by + 0.5
- Whenever a midfielder is substituted instead of a forward, the measure will be decreased by - 0.5
- Whenever a defender is substituted instead of a non-defender, the measure will be decreased by - 1
- Whenever a neutral (e.g. forward instead of forward) substitution is made, the measure will retain its current value

Thus, over the course of a match, depending on the type of substitutions, a team's differential represents the offensiveness of the strategy adjustments. In Figure 5, the average of all matches' substitution differentials is presented in a graph per playing minute.

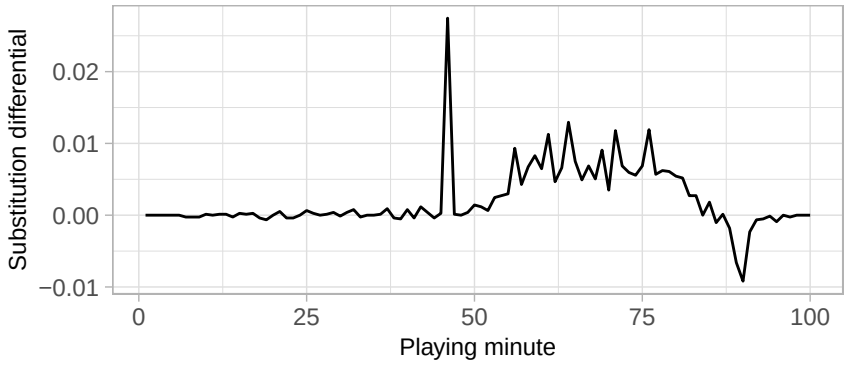


Figure 5: Substitution differential per playing minute

Figure 5 shows the average of the substitution differential measure per minute over the course of a match. It can be noticed that, on average, coaches do not make strategy adjustments in the first half of a match. It seems that offensive substitutions are most likely to be made right after the break,

which is a natural moment for coaches to make substitutions. During the second half, a preference for offensive substitutes can be observed as well. As the end of the match approaches, a tendency arises towards more neutral and defensive substitutions. The average values of the substitution differential measure are 0.422 on a match level and 0.002 on a minute and team level. This implies that, on average, coaches tend to implement a more offensive substitution strategy over the course of a match.

3.5 Division in training and test set

Within this research, the dataset will be split into a training and test set. The training and test sets are randomly drawn, and consist of 80% and 20% of the complete dataset respectively. To make a fair comparison between models' prediction accuracy, the data points in the test set remain unseen in the training phase of the model. The test set will only be used to check all models' performance.

When tuning the different machine learning model's parameters, Cross-Validation will be performed on the training dataset. Hereby, it is possible to approach or reach the most optimal tuning parameters for each respective model. Cross-Validation is a technique where the training set is split randomly in a number of K folds. In this research, we use $K = 10$, which is a common value to use. After the splitting of the data, the model will be trained on the data of $K - 1$ folds of the data and validated on the K^{th} fold that is left out. This process will be repeated until all folds have been used as a validation set. The average prediction accuracy over all folds can be computed, and this will provide information about the set tuning parameter. The process can be repeated for different values for the model's respective tuning parameter, whereafter the optimal tuning parameter will be found where the prediction accuracy was highest. Hereby, the test set is left out of the tuning process, and remains unseen, so that it can be used to compare the predictive performance of different models later on in this research.

3.6 Limitations of data set

The most important limitation of the data set is that no other field positions of all players other than the player in possession of the ball are described. For example, the 'XG' metric (explained in Section 4.2) which desribes the quality of chances a team has had during a match, could be more advanced when the field positions of the opponent's defending players are known. Namely, in that case it would be possible to evaluate the space that the attacking player has at the moment of the

goal-attempt, and thereby the quality of a chance can be adjusted accordingly.

Furthermore, when a player dribbles with the ball, it is not described as event in the data set, so no information is known about a team's dribbling abilities.

Also, individual player data, which for each player describes its total running distance per match, maximum sprinting speed and its high-intensity running (HIR) periods, would provide a new dimension of performance data where the team's overall work-rate, as well as a player's individual work-rate, could be researched. As discussed in Section 2.2, the HIR variable appears to be a comprehensive metric to describe players' physical performance.

4. Football match statistics

From the retrieved data set it is feasible to extract a set of predictor variables that might be relevant when trying to classify the match outcomes. At first, various basic football metrics are explained, followed by more advanced metrics later this section.

4.1 Basic match statistics

The following basic football metric variables are extracted from the retrieved data set for every match.

- **Ball possession (%)**: a percentage figure of the total amount of time a team has possession of the ball.
- **Passes**: the total number of passes a team makes, both accurate and inaccurate.
- **Pass accuracy (%)**: a percentage figure of the amount of accurate passes opposed to the total number of passes made by a team.
- **Shots**: the total number of shots a team makes, both on target and off target.
- **Shots on target**: the total number of shots on target a team makes. A shot is defined as 'on target' when the ball goes into the net regardless of the intent, or when the ball would have gone into the net but the goalkeeper or the last-man player blocked the goal attempt. In this research, a shot that directly hits the post or the crossbar is also defined as 'on target'.

- **Fouls**: the total number of fouls a team makes. A foul occurs when the referee deems a player to make an unfair act, that violates the rules of the game. When a foul is made, the opposing team will be awarded a free kick or a penalty (when the foul is made within the box close to its own goal).

- **Yellow cards**: the total number of yellow cards shown to a team. A yellow card is shown by the referee when the player makes a significant foul, which in the referee's eyes deserves an official caution. When a player receives two yellow cards, he must leave the pitch and the team is not allowed to replace the player by a substitute, forcing their team to play with one player fewer.

- **Red cards**: the total number of red cards shown to a team. A red card is shown by the referee when a player commits a serious offense, such as a violent foul. The respective player is obliged to leave the pitch after receiving the red card. The player may not be replaced by a substitute, forcing their team to play with a player fewer.

- **Duels won**: in a game duels occur between players of opposing teams who both aim to win ball possession for their team. Such duels can be: air duels, ground duels or loose ball duels. This metric describes the total number of duels won by a team.

- **Crosses**: the total number of crosses a team makes. A cross is a pass (medium- to long-range) from the wide area of the pitch towards a central position of the pitch close to the opponent's goal. The cross can be made from an active play, but also a restarting play (e.g. free-kick or corner).

- **Corners**: the total number of corners a team earns. A corner is awarded to a team if the ball crosses the goal line and the final ball touch is made by the defending team. In essence, a corner is a free kick that has to be taken from the corner of the pitch located in extension of the previously crossed goal line.

- **Free kicks**: the total number of free kicks a team earns. When the opposing team commits a foul, a free kick is awarded at the same position where the foul is made.

- **Penalties**: the total number of penalties a team earns. When the opposing team commits a foul inside its own box (near its own goal), a penalty will be rewarded. A penalty kick is

taken from the penalty spot, which is located 11 meters away from the goal. No opposing players are allowed to block the penalty kick except the goalkeeper, which makes it a great goal-scoring opportunity.

4.2 Expected goals

The **Expected Goals (XG)** metric is a more advanced metric than the ones mentioned before. The metric represents the quality of a goal attempt based on the type of shot and the position where the shot was taken from. Thereby, adding up the team's 'XG' metric indicates how many goals a team should score on average based on the quality of the shots the team takes. In this research three types of shots are distinguished: headed shots, normal shots and penalties. Each type has its unique goal-scoring probability per field position of the shot. It turns out that the parameters 'distance to the goal' and the 'angle to the goal' from where the shot is taken have a huge influence on the chance of a goal attempt finding the net. These probabilities for each shot type are based on the fraction of all goals scored from that specific field position over all attempts made from the position for the same shot type. Unfortunately, the data set is limited to only the event position of the player who has possession of the ball. Having information about the other players' positions further improves the 'XG' metric, because then it is known whether a shot is a 'free shot', or if a great part of the angle to the goal is blocked by an opposing player. The quality of the shots can thus be better evaluated. A key point to mention is that the player's skills are all neutralized by this metric, since it averages over all goals scored by all players in all leagues in the data set. The skills of a forward to score from an almost impossible position will not be taken into account in the 'XG' metric. This might lead to the situation where a team's 'XG' metric for a given match is low, but the amount of goals scored is much higher. Also the opposite could hold, where a team's 'XG' metric is high, but the amount of scored goals is low. This indicates that the team is underpeforming on the finishing of its chances.

The core of this thesis will be to investigate the effect of substitutes on important football metrics. The 'XG' metric does give a good general indication about the quality of chances a team has, and the influence a substitute has on the metric can be interesting. Later in this research, it will be further evaluated.

In Figures 6 and 7 heatmaps are presented of all normal shot and heading shot pitch positions occuring in the training data set. The direction of attack is from left to right, where the opponent's

goal is located in the centre on the plot's right side. A shot is defined as 'normal shot' if the shot in `events data` was made by the player's feet and not with his head, and if it was not a penalty shot. Thereby, it consists of shots made in regular plays and also free kick shots. A shot is defined as 'heading shot' if the shot in `events data` is made by the player's head, which is only possible in active plays. All individual points in Figures 6 and 7 indicate the probability of the shot from that position resulting in a goal. The probability is calculated by the fraction of goals scored from that position over all shots made from that position. All occuring shots in the training set are used for these computations. In the left plot, note that some position have a probability of 1.0, because only one shot has been made from the position that also resulted in a goal. If we do not adjust this, it will lead to false interpretations by the model later on. Therefore, in this research, only positions where at least three shots were taken from in the training set of the 2017-2018 season in all seven leagues, are incorporated in the 'XG' metric. This results in the right plots in both Figures 6 and 7.

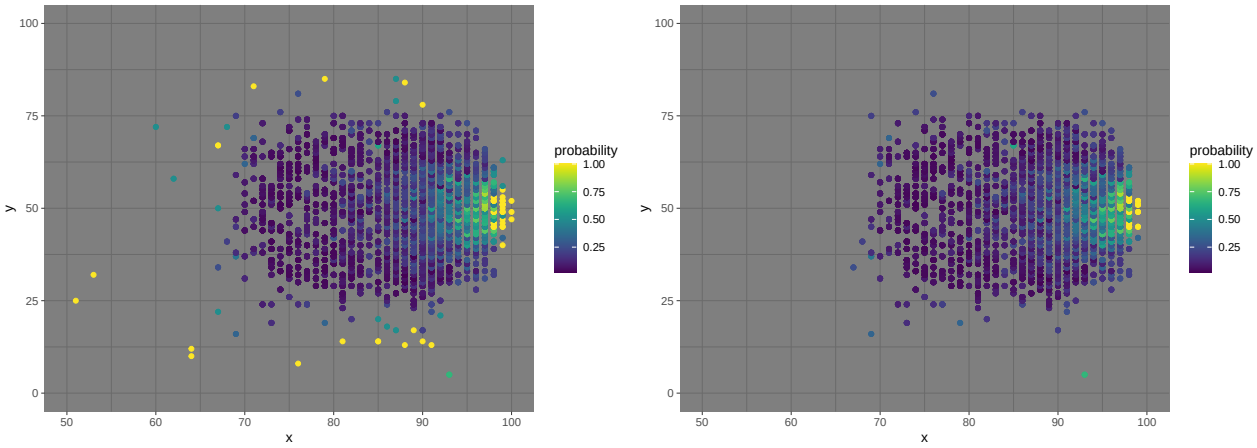


Figure 6: Normal shot goal-scoring probabilities per position

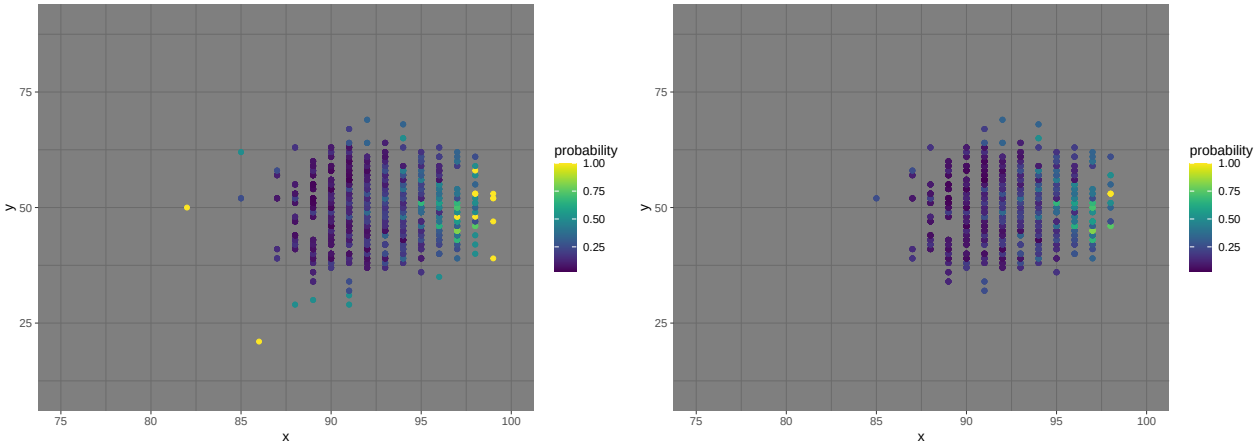


Figure 7: Heading shot goal-scoring probabilities per position

The x-axis ranges from $[50, 100]$ in Figure 6, and ranges from $[75, 100]$ in Figure 7. It is observable that header goals are scored from a much closer range to the opponent's goal than normal goals, which is very intuitive when considering the potential ball speed of both type of shots.

Additionally, the probability of a penalty shot finding the net will be calculated as the fraction of all scored penalties over all penalties taken in the training dataset. Note that each shot type (header, normal shot, penalty) is now accompanied by its unique goal-scoring probability distribution. Finally, per match the 'XG' metric for team i will be computed by the sum of the goal-scoring probabilities of each team's shot, distinguishing the type of shots between normal shots, heading shots, and penalty shots.

4.3 Invasion index

Another more advanced metric is the **Invasion index**, which is proposed by **Pappalardo et al. (2019)**. It is a measure of how close to the opponent's goal a team plays on average. It indicates the *dangerousness* of a team. Since the data set comprises the positions of all football match events occuring in a match, it is possible to compute the invasion index for each possession phase of a team. The possession phase is a sequence of consecutive events on the ball made by a team, before the opposing team gains possession or the ball goes out of play. In this research, a sequence of events is defined as a possession phase when a team has possession of the ball until the sequence is interrupted. A sequence is interrupted when the ball is out of play, when the ball possession holding team loses a duel, or when a foul is made by the opposing team. Thereafter, a new possession phase will be started.

Finally, the invasion index is computed by evaluating all field positions of a possession phase. For each reached field position, the goal scoring probabilities are calculated as the fraction of goals scored from that position within the complete data set (all matches in seven leagues). Eventually, the invasion index is defined as the highest goal-scoring probability of all events' positions within a possession phase. The overall invasion index of a team during a match is computed by the average invasion index over all the team's possession phases. Note that the invasion index does not distinguish between types of plays when calculating the positions' goal-scoring probabilities, as the 'XG' metric does.

Generally, a high invasion index indicates a team is playing close to the opponent's goal, and logically

would reflect the team's performance. In later sections of this thesis the importance of the invasion index is analyzed in combination with the substitutes' impact on respective metric.

4.4 Acceleration index

The **Acceleration index** is a metric based on the invasion index, and is also proposed by **Pappalardo et al. (2019)**. It measures how fast a team reaches the most dangerous event of a possession phase. Recall that the most dangerous event per possession sequence is comprised by the invasion index. The computation of the acceleration index is the ratio between the first event and the most dangerous event of the possession phase. To calculate a team's overall acceleration index during a match, just like the invasion index, the average acceleration index is calculated over all the team's possession phases.

Thus, the acceleration index is also a measure of how dangerous and quick a team is as regards the creation of goal-scoring opportunities. The metric indicates a team's playing efficacy during a match. In later sections, the importance of respective metric within the classification problem on hand will be analyzed.

4.5 Passing indicator

The **Passing indicator** metric is a football metric proposed by **Pappalardo et al. (2015)**. The authors claim that the respective indicator has a clear correlation with the total goals scored, total goal attempts, and points obtained in the final rankings. By the authors, the passing activity of a team is claimed to have a relationship with the team's success. The passing indicator is composed of five individual metrics which are combined into one. The five individual metrics are the total passing volume, ω, the mean players' passing volume, μ_p, the variance of players' passing volume, σ_p, the mean zones' passing volume, μ_z, and the variance of the zones' passing volume, σ_z. To perform the partitioning of the pitch into equally sized zones, $10 \times 10 = 100$ zones are created. There exists differentiation amongst the exact pitch sizes per football stadium, but the most preferred pitch size amongst professional clubs is 105 by 68 metres. Based on this standard size, the zones created for this research each have approximately a size of 10.5 by 6.8 metres.

After having computed all individual metrics, the passing indicator, H, is computed as the harmonic mean of the latter.

$$H = \frac{5}{1/\omega + 1/\mu_p + 1/\sigma_p + 1/\mu_z + 1/\sigma_z}$$

The measured importance of the passing indicator within the classification problem is reviewed later within this thesis.

4.6 Defensive metric (PPDA)

Since coaches can make both offensive and defensive substitutes, depending on the coach's intention and scoreline, this research also evaluates defensive metrics. A well-known defensive metric is the **Passes Allowed Per Defensive Action (PPDA)**, which measures the intensity of a high pressing defensive team. Defensive actions are defined as duels (won or lost), fouls, and interceptions. The metric only evaluates passes and defensive actions made at the opponent's 60% of the pitch (calculated from the opponent's goal).

However, the PPDA does not display the quality of the pressing, which makes interpretation more difficult and some caution is necessary. Nevertheless, dominant teams usually have lower a PPDA because of their high pressing defensive play. The importance of this metric will also be analyzed later on in this thesis.

5. Methodology

As mentioned in earlier sections, the first part of this research is to find the key team performance indicators. To achieve this, some supervised machine learning methods will be implemented to classify the match outcomes. In this section, an explanation of the used methodology will be provided.

5.1 Classification methods

This thesis includes the fitting of four different classification models on the training set. Hereafter, all models' predictive performance will be shown by predicting the (unseen) test set data. Thereby, a comparison between models' performance can be made and only the variable importance measure of the best-performing model will be further evaluated. By such evaluation, the **first research question** can be answered.

5.1.1 Ordinal logistic regression

The Ordinal Logistic Regression (OLR) is an extension to the binary logistic regression. When performing a binary logistic regression the response variable is categorical and comprises two levels. In this research the categorical variable can take on three levels: a team can win, lose or draw. Besides that, the dependent variable comprises more than two categories and the categories are ordered. In cases of explicit ordering in the categories of the dependent variable Y, OLR can be used to handle such a multi-class classification problem. OLR is a generalized linear model used to estimate the probabilities for the J categories of ordinal outcome Y, with the use of a set of predictor variables X. To compute the final outcome the OLR will fit $J-1$ independent binary logistic models. This means that one of the categories of the response variable must be defined as a reference. All other categories will be compared to the respective reference category. It can be defined as follows:

$$\frac{P(Y \leq j)}{P(Y > j)}$$

for $j = 1, ..., J-1$. The *log odds*, also known as *logit*, are computed in cumulative probabilities:

$$log\frac{P(Y \leq j)}{P(Y > j)} = logit(P(Y \leq j))$$

The OLR model is parameterized as:

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - ... - \eta_p x_p$$

where there will be $J-1$ equations in total because the probability of being in at least the highest category is 1. To calculate the logit, the model estimate will be subtracted. The resulting values represent cumulative probabilities.

The OLR model holds onto the proportional odds assumption. To fulfil the proportional odds assumption, the logarithms of odds (not the probabilities) must form an arithmetic series. The final classification of a test observation is done by checking which category probability is highest.

The important output by the OLR model is the logistic coefficients. If all other variables remain

equal, a predictor variable's coefficient represents the expected amount of change in the *log-odds* for each one-unit change in the respective predictor variable. The model allows p-values to be calculated for each coefficient, indicating whether the coefficient is statistically significantly different from zero or not. Thus, by the output of OLR the most important variables, including the magnitude of their impact on the reference category of the dependent variable, can be derived.

5.1.2 Random forest

Since regular decision trees are very prone to overfitting they usually have very high variance. The random forest method aims to reduce the variance and overfitting issue that regular decision trees are facing. The random forest method builds many regular decision trees as an ensemble and then makes a final prediction on the collection of them. Random forests are known to have a much better predictive accuracy than a single decision tree. A random forest is built on the principle of bootstrapped training samples. This is a technique where slightly different training samples are drawn. These bootstrapped samples are drawn with replacement, which results in each sample being slightly different from the original training data set. Hence, the grown trees of the random forest are each built on a different training sample. Furthermore, at every split in each tree, only a random sample of m predictors will be allowed to be evaluated as predictor variables. Typically, for classification problems, $m \approx \sqrt{p}$ (rounded down), where p is the total number of predictors. However, m must be treated as a tuning parameter and it depends on the problem what is the optimal m that leads to the best predictive performance of the model. Since random forests only use a subset of predictor variables at each node, it brings an extra element of randomness to the model. Thereby, the grown (unpruned) trees will each be a little bit different structure-wise. In general, this corresponds to a decorrelation between all grown trees. Finally, the classification of an observation is done by the majority vote of the resulting predictions over the collection of unpruned trees. This will reduce the model's variance by a great amount.

At each split in the trees, the Gini index is considered to determine which variable can make the best split in terms of classifying the data. The Gini index (G) is defined by:

$$G = \sum_{j=1}^{J} \hat{p}_{mj}(1 - \hat{p}_{mj}) \tag{1}$$

In this equation, $\hat{p}_{mj}$ represents the proportion of training observations from the jth class in the

mth region. Thus, the Gini index is a measure of *node purity* **(James et al., 2017)**. The smaller the value of G, the more a node contains predominantly observations from a single class. Thus, the Gini index will be minimized by the algorithm at each splitting point in the tree.

The optimal tuning parameter m (mtry) can be found by using the technique K-Fold Cross-Validation on the training set.

Finally, this research aims to detect the most important variables within the process of classifying the data. To do so, the total decrease in Gini index caused by using a given predictor variable must be computed, and then be averaged over all splits in the forest involving the predictor variable in question. The higher the 'Mean decrease in Gini' value, the more important the role of the predictor variable is within the classification process. An overall summary can be obtained for all predictor variables whereby the corresponding variables' importance can be compared.

5.1.3 Support vector machines

The Support Vector Machine (SVM) is a classifier that is based on the construction of separating hyperplanes. SVMs are able to classify classes separable both by linear and non-linear class boundaries. A hyperplane classifier is defined by the drawing of separating lines which classify members of different classes.

The width of the parallel space to the separating hyperplane, where no interior data points are located, is called the margin. When the margin between the separating hyperplane and the classes is greatest, the optimal separating hyperplane is found for the classification problem. Usually, the class boundaries are complex shaped in order to find the optimal classification of the data points (non-linear). In such cases, the SVM approach is to transform the data points to a higher dimension prior to finding the optimal separating hyperplane, where the data points are hopefully linearly separable. The transformation of the original data points is done using mathematical functions that are called "kernels". A popular kernel used by SVMs to tackle non-linear classification problems is the 'radial' kernel. When the SVM kernel is set to 'linear', the data will not be transformed to a higher dimension. The radial on two different samples x and y, represented as feature vectors in some input space is defined as:

$$k(x, y) = \exp(-\gamma \|x - y\|^2)$$

where γ is the tuning parameter that controls the variance of the model and accounts for the smoothness. Furthermore, $\|x - y\|^2$ is recognized as the squared Euclidean distance between the two samples. For very large γ, the decision boundary is quite fluctuating, which might lead to overfitting. For smaller values of γ, the decision boundary has lower variance and is smoother. In essence, γ is scaling the influence. When points are relatively far from each other, this equation will go towards zero. Thus, the further two observations are from each other, the less influence they have on each other. When plugging in the values of observations, we get the high-dimensional relationship value. Because the radial kernel finds SVMs in infinite dimensions, it is not possible to visualize what it does.

By its nature, SVM does not lend itself to classify more than two classes. However, extentions such as the *one-versus-one* and *one-versus-all* approaches allow the SVM to be applied to multi-class classification problems as well. In this research the *one-versus-one* approach will be applied to the classification problem. During this approach, all possible combinations of pairs of classes K are computed, which can be denoted by $\binom{K}{2}$. Then, for each pair, a binary SVM is developed. Subsequently, a test observation is classified using each of the $\binom{K}{2}$ classifiers, and the class is assigned where the test observation is assigned to most.

When performing a SVM with radial kernel, the optimal tuning parameters for cost (C) and γ are found by using a grid search and K-Fold Cross-Validation on the training set. However, when the linear kernel is implemented, only C has to be tuned. Both the linear and radial kernel will be used by the SVM model later in this research.

5.1.4 Naive Bayes

In contrast with previously discussed machine learning models, the Naive Bayes classifier has a probabilistic approach. It is founded on the concept *conditional probability*, which is defined as the likelihood of the occurrance of event A given that event B occured, denoted as $p(A|B)$. The model relies on the very strong, and usually unrealistic, assumption of independence between predictor variables. This implies that there is no correlation between these features, which explains the algorithm's name "naive". The core of the classifier is Bayes' theorem. The theorem is defined in Equation (2), where $p(A)$ and $p(B)$ represent the probability of individual events A and B occuring.

$$p(A|B) = \frac{p(A)p(B|A)}{p(B)} \tag{2}$$

Given data point $x = \{x_1, ..., x_n\}$ of n features, for $j = 1, ...J$ the Naive Bayes algorithm is able to predict class C_j given x according to the probability

$$p(C_j|x) = p(C_j|x_1, ..., x_n)$$

By using Bayes' Theorem it can be factored as

$$p(C_j|x) = \frac{p(x|C_j)p(C_j)}{p(x)} = \frac{p(x_1, ..., x_n|C_j)p(C_j)}{p(x_1, ..., x_n)}$$

Then, by the "naive" independence amongst features assumption, it can be further decomposed into:

$$p(x_1, ..., x_n|C_j) = p(x_1|C_j)...p(x_n|C_j)$$

Thus,

$$p(C_j|x_1, ..., x_n) \sim p(C_j) \prod_{i=1}^{n} p(x_i|C_j)$$

The final classification involves assigning class C_j to the observation for which $p(C_j|x_1, ..., x_n)$ is greatest. Specifically when, $p(C_a) \prod_{i=1}^{n} p(x_i|C_a) > p(C_b) \prod_{i=1}^{n} p(x_i|C_b)$. Thus, estimated class $\hat{C}$ is defined as

$$\hat{C} = \underset{j \in \{1,...,J\}}{\operatorname{argmax}} p(C_j) \prod_{i=1}^{n} p(x_i|C_j)$$

The Naive Bayes classifier is known to work well with high-dimensional data and is computationally very fast. However, due to the strong assumption of independence, the algorithm is usually less accurate than more complicated machine learning algorithms. The tuning parameters for the Naive Bayes model are *Laplace smoothing* and the option to include kernel density estimation (KDE).

Laplace smoothing addresses the problem of missing values such that the count of each feature is always non-zero. Namely, if some feature would have zero counts, the probability estimate would be zero. Laplace smoothing adds a small value, or pseudo-count, α, to the count of each feature, such

that all probability estimates for any combination of features will always be non-zero. The Laplace smoothing technique is primarly used in classification problems where the feature counts may be sparse.

The other tuning parameter, KDE, is a way to estimate the probability density function of a random numeric variable. When including continuous predictor variables normality is often assumed, such that the probability can be computed from the variable's probability density function. However, this normality assumption is not always fair. Therefore, the probability distribution can be derived by KDE. Thereby, the aim is to get a more accurate representation of continuous variable probabilities. KDE is essentially a non-parametric estimator of density.

Within KDE, the kernel bandwidth determines the size of the kernel at each point, where a larger bandwidth means a more flexible density estimate. When adjusting the bandwidth parameter, the bias-variance trade-off applies. Because, low bandwidth values lead to high-variance estimates (overfitting), whereas high bandwidth values lead to high-bias estimates (under-fitting). This tuning parameter will be tuned by the Cross-Validation method.

5.2 Partial dependence plots

In order to answer the **second to fifth research question**, the relationships between the important predictor variables and the classification of the different match outcomes (lose, draw, win) must be visualized. Such understanding is relevant when the substitution's impact is being measured on those variables in a later stage.

Machine learning models may be very good in classifying data, but can also be regarded as "Black Box models", where it is difficult to interpret its results. A Partial Dependence Plot (PDP) is a very useful tool for detecting the direction of a given predictor variable's effect on the response variable and is able to reveal insights on the model's output. By the PDP, the relationship between the response variable and given predictor variable can be visualized in a plot. In these plots the average impact on the class probability of the response variable is presented for a range of values for a given predictor variable. By the PDP, the individual marginal effect of the predictor variable on the model outcome is illustrated. The remaining predictor variables are ignored within this process. Hence, only when a given predictor variable is uncorrelated to any of the remaining predictor variables, it will perfectly reflect the true marginal effect on the average model outcome. Logically,

in practice, it is uncommon that a given predictor variable will not have any interaction with other predictor variables. Thus, the PDP must be cautiously treated in order to analyze predictor variables' marginal effects. Knowing there is some probability for error, this research only uses the PDP to globally analyze the effects, and will not dive into too much detail.

From the model's perspective, the partial plots denote the relative contribution of the given predictor variable on the given outcome class probability. Having said that, negative values in the PDP on the y-axis imply that the given outcome class is less likely for respective value of predictor variable (x-axis). Similarly, the opposite is also true, where positive values on the y-axis imply a higher probability for the given outcome class for those values of the predictor variable. Zero values on the y-axis imply no average impact on the probability of the outcome class. Note that it is important to check how the partial plot is defined before inferring any conclusions.

In the Appendix the most relevant PDPs for this research are represented.

5.3 Panel data analysis

The final stage of the analysis consists of creating a dataframe that represents the real-time match statistics in 3 minute intervals for a subset of N matches. The **second to fifth research question** will be answered by the implementation of a 'substitution dummy' variable. In other words, this research wishes to investigate the effect and significance of the created dummy variable that indicates whether a team has offensively or defensively substituted at time t or not. The previously found important match statistic variable (found by the first research question), on which the substitutes' impact must be evaluated, will be the dependent variable, and several other important match statistic variables (including the 'substitution dummy' variable) are used as predictor variables. Finally, all regression coefficients can be evaluated, including the respective p-values which indicate the significance of each independent variable regarding the ability to predict the outcome of the dependent variable. Hence, to answer the **second to fifth research question** the significance and magnitude of the coefficient of the 'substitution dummy' variable will be evaluated.

The substitutions' impact will be reviewed under different scoreline scenarios, and both offensive and defensive substitutions are researched. The subset of investigatable N matches are matches where offensive or defensive substitutions have been made by a given team, depending on which research question must be answered. The dataframe that will be created is panel data since it is

longitudinal ($T = 30$) and cross-sectional (over N entities/matches). Because the data is collected over time and the same entities, a regression can be run over these two dimensions. In essence, a panel data regression model looks like:

$$y_{it} = a + bx_{it} + \epsilon_{it}$$

where y represents the dependent variable, x the independent variable, i and t the indices for entities and time, and a and b the coefficients. Panel data analysis has two approaches, depending on the assumptions made about the error term ϵ. These two approaches can be distinguished into a fixed effects or a random effects model. The difference is that in a fixed effects model, ϵ is assumed to vary non-stochastically over i or t, whereas in a random effects model ϵ is assumed to vary stochastically over i or t.

In other words, when using the fixed effects model, it is assumed that something within each individual entity may impact the predictor variable and we need to control for this. Another assumption of the fixed effects model is that each individual entity's time-invariant characteristics are not correlated with other individual entity's characteristics. Hence, each entity is assumed to be different, and therefore the entity's error term and the constant (that captures the individual characteristics) should not be correlated with the other entities. When the error terms are correlated between individual entities, the fixed effects model may not be suitable, and the random effects model may be preferred. When using a random effects model, the variation across entities is assumed to be random and uncorrelated with the model's predictor variables. Thus, when there is reason to believe that individual entities have different influences on the dependent variable, the random effect model should be used.

The Durbin-Wu Hausman test can be used to test whether a fixed effects model, rather than a random effects model, should be used. The null hypothesis of this test is that the preferred model is random effects versus the alternative hypothesis that the preferred model is fixed effects. After having selected the appropriate model, the regression can be run on the panel data, and the coefficient along with the significance of the 'substitution dummy' will be inspected.

5.3.1 Different types of dependent variables

Within this research, the dependent variable used in the panel data analysis can either be a continuous or a discrete (count) variable.

- When the dependent variable is continuous (e.g. Expected goals (XG)), a linear regression model will be performed.
- When the dependent variable is a count variable (e.g. number of duels won), a Poisson regression model is performed.

5.4 Mann-Whitney U test

In order the answer the **sixth research question**, values for two groups must be compared statistically. If two groups are assumed to be normally distributed, the Student's t-test is a well-known test to use. However, the assumption of a normal distribution does not necessarily hold. In cases when this crucial assumption does not hold or is suspected to fall short, the Mann-Whitney U test is a preferred test that helps to answer such problem. Like the Student's t-test, the Mann-Whitney U test provides results suggesting whether the values of two groups are statistically different. The test can also be performed when the data on hand is unpaired, meaning that the two groups differ in size of total observations. In Section 6.4 this test will be performed to answer the last research question.

6. Results and Analysis

6.1 Comparing models' performance

Within this section the performed models are compared based on their predictive performance on the (unseen) test set data. In Table 7, the prediction accuracies are given.

Table 7: Predictive performance of all models on test set

Classifier	Accuracy
OLR	56%
Random forest	57%
SVM (linear kernel)	56%
SVM (radial kernel)	58%
Naive Bayes	51%

Note:
Prior to this, all models' hyperparameters were tuned and set to the optimal performing parameters by the Cross-Validation method

By the presented results in Table 7, it is observable that the OLR performs almost as well as the Random forest and SVM methods. By evaluating the OLR confusion matrix (see Appendix), it can be observed that the OLR does not predict any test observation's class to be of category 'Draw'. Thus, the estimated probabilities which correspond to the class category outcome 'Draw' are always lower than the estimated probabilities for the categories 'Lose' and 'Win'. On the other hand, the Random forest and SVM methods are predicting draws, and both perform slightly better than the OLR method.

For this reason, the variable importance gained by the Random forest or SVM will be reviewed instead of the OLR coefficient estimates. In the Appendix, Tables 14-16 represent the corresponding confusion matrices of the OLR, RF, and SVM models.

6.2 Variable importance and PDP

In previous section the models' prediction performance on the test set were given. Thereby, it can be concluded that the Random forest and SVM method are the best predicting classification models for the problem on hand. By its nature, SVM is a model that is known for its moderate interpretability. On the other hand, Random forest's interpretability is reasonably well. In Section 5.1.2 it was explained how the variables' importances are calculated by the Random forest model. In Figure 8, the variable importance plot by the Random forest model is presented.

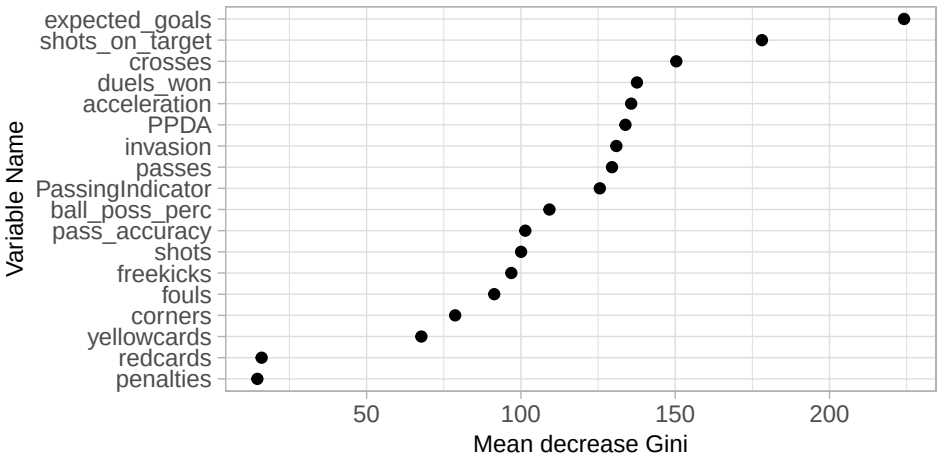


Figure 8: Variable importance by Random forest

From Figure 8, it can be clearly observed that the variables 'Expected goals (XG)' and 'Shots on target' are considered as the most important variables within the classifiying task. These two variables are measuring the offensive aspects contributing to team performance. Meanwhile, the variables 'Duels won' and 'PPDA' are the highest ranked variables as regards to the defensive aspects contributing to team performance. Hereby, an answer is provided for the **first research question** on hand.

Continuously, in order to answer the **second to fifth research question** on hand, the average substitutions' impact on these four mentioned variables will be evaluated. But prior to analyzing the substitutions' impact, the PDPs, which are plots that indicate the effect of respective predictor variables on the response variable, must be evaluated. In Figure 9 in the Appendix the four PDPs are plotted on the match outcome 'Win'.

Regarding the offensive metrics, it can be observed that a higher 'XG' metric generally leads to an increase in a team's chances to win a match. Similarly, this holds for the offensive metric 'Shots on target'. Both insights are intuitively correct, because a rise in a team's total goal attempts is likely to lead to more balls eventually finding the net.

Regarding the defensive metrics, a rise in the metric 'Duels won' is likely to increase a team's chances to win a match. However, as can be observed, the effect is smaller than the relative increase gained by increasing one of the offensive metrics. Concerning the partial dependence on the 'PPDA' metric, it can be stated that all values for PPDA seem to have a positive contribution on the chances of

winning, which makes it redundant in predicting wins. It can be concluded that the 'PPDA' metric does a better job in predicting draws as opposed to winning and losing, and therefore the metric does not contribute to the purpose of this research. Therefore, no further research will be done regarding the substitutions' impact on the 'PPDA' metric, because it will not lead to insightful results.

6.3 Panel data regression results

In this section the results of the panel data analysis will be presented. As mentioned before, the type of panel data regression depends on the type of the dependent variable (discrete or continuous). To start with, different scoreline scenarios are delineated along with different substitution strategies. To answer the **second, third and fourth research question**, the following different scenarios are considered within the panel data analysis:

- Team is *drawing*, and the coach substituted *offensively*;
- Team is *behind*, and the coach substituted *offensively*;
- Team is *drawing*, and the coach substituted *defensively*;
- Team is *ahead*, and the coach substituted *defensively*.

Note again that, within this entire research, offensive or defensive substituting is defined as the situation when a coach substitutes *more* offensively or defensively than its opponent. If the two opposing teams substitute equally offensively or equally defensively, the respective match data will never be incorporated into any analysis, because the substitute differential (discussed in Section 3.4) will be equal to zero meaning that no insightful conclusions can be inferred. Within the analysis only match data is considered where such a scenario occurred. These regressions' results can be found within this section in Tables 8-10.

To answer the **fifth research question**, covering the difference between top-ranked and low-ranked teams, the scenarios where top-ranked and low-ranked teams held onto an *offensive* or *defensive* substitution strategy will both be considered and compared. Thus, the following scenarios (also along with different scoreline scenarios):

- The coach of a top-ranked team substituted *offensively*;
- The coach of a low-ranked team substituted *offensively*;
- The coach of a top-ranked team substituted *defensively*;

- The coach of a low-ranked team substituted *defensively*.

The summaries of performed regressions are presented within this section in Tables 11-13.

Note that, when interpreting the results, there exists a distinction between the performed regression models. When performing a Poisson regression model, the expected value of the dependent variable, E(y), is related to the independent variables by the log link function. Which can be formulated as follows:

$$\mathrm{E(y)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$$

Therefore, the coefficients of the Poisson regression have to be interpreted differently compared to a regular linear regression. By the Poisson regression, a one-unit increase in the independent variable x_n yields a factor equal to $\exp(\beta_n)$, which multiplied by the dependent variable, indicates the expected value of the dependent variable due to a one-unit increase in x_n.

The coefficients of the linear regression model can be interpreted as follows: a one-unit increase in the independent variable x_n yields an expected increase of the dependent variable equal to the value of respective coefficient, β_n.

Furthermore, related to each performed regression, the Durbin-Wu Hausman test determines whether a fixed or random effects model should be performed. The obtained results will be presented in this section, and the drawn conclusions and more detailed answers to all research questions will be provided in Section 7.1.

The results presented by Tables 8-10 provide answers to the **second, third and fourth research question** and can be interpreted as follows. From Tables 8 and 10, it can be derived that when a team is drawing, and when the coach adjusts its team's strategy either offensively or defensively, the impact on 'XG' and 'Duels won' is generally smaller than the impact in cases where the team is behind or ahead. Actually, the substitutes' impact on 'XG' when drawing is insignificant, in contrary to the significant positive impact on 'XG' when teams are behind. The average impact of offensively substituting, when team are behind and keeping all other independent variables' values equal, is 0.01. This means that in general, an offensive substitution strategy will increase a team's expected goals by 0.01.

Table 8: Panel data regression where 'XG' is the dependent variable (offensive substituting)

	Coefficients (p-values)	
Variables	**Drawing**	**Behind**
Substitution_dummy	-3e-03	0.01**
Shots_on_target	0.12**	0.11**
Duels_won	1e-03*	2e-03**
PPDA	-6e-04**	-7e-04**
Invasion	0.57**	0.43**
Acceleration	-2e-03	-2e-03
Crosses	4e-03**	5e-03**
Passes	-6e-04**	-6e-04**

Note:
The columns 'Drawing' and 'Behind' denote the scoreline scenario at the moment of the offensive substitution; (*p < 0.05, **p < 0.01)

Table 9: Panel data regression (Poisson) where 'Shots on target' is the dependent variable (offensive substituting)

	Coefficients (p-values)	
Variables	**Drawing**	**Behind**
Substitution_dummy	0.19**	0.11*
XG	3.17**	3.52**
Duels_won	0.09**	0.08**
PPDA	-0.03**	-0.02**
Invasion	2.38**	3.91**
Acceleration	-0.05	-0.21**
Crosses	0.09**	0.03
Passes	2e-03	8e-03**

Note:
The columns 'Drawing' and 'Behind' denote the scoreline scenario at the moment of the offensive substitution; (*p < 0.05, **p < 0.01)

The other important offensive metric to analyze is a team's total amount of shots on target. In Table 9 it can be observed that in cases of drawing or being behind, the offensive substitutions have a positive impact on the 'Shots on target' metric. In cases of drawing, the average impact seems to be even bigger than when teams are behind. When drawing, the average substitutes' impact on the shots on target is equal to $\exp(0.19) = 1.21$ (a 21% increase), whereas the impact when behind is equal to $\exp(0.11) = 1.12$ (a 12% increase). Note again, that in both cases all other independent variables' values must be kept equal for such interpretation.

The results in Table 10 show that, in general, when teams are ahead, the defensive substitutes' impact on the total amount of duels won is bigger than when drawing. In both scenarios the defensive substitutes' average impact is positive and statistically significant. The coefficients of the Poisson regression model should be interpreted as $\exp(0.07) = 1.08$ and $\exp(0.09) = 1.10$, indicating an 8% and 10% average increase on the total amount of duels won, dependent on the scoreline.

Table 10: Panel data regression (Poisson) where 'Duels won' is the dependent variable (defensive substituting)

	Coefficients (p-values)	
Variables	**Drawing**	**Ahead**
Substitution_dummy	0.07**	0.09**
Shots_on_target	0.07**	0.07**
XG	0.15*	0.02
PPDA	-0.01**	-0.01**
Invasion	0.29	0.08
Acceleration	0.04*	0.02*
Crosses	0.07**	0.10**
Passes	0.01**	0.01**

Note:
The columns 'Drawing' and 'Ahead' denote the scoreline scenario at the moment of the defensive substitution; (*p < 0.05, **p < 0.01)

The results presented by Tables 11-13 provide answers to the **fifth research question** and can be interpreted as follows. When teams are behind, offensive substitutes of top-ranked teams seem to have a statistically significant positive impact on the 'XG' metric. The same holds for offensive subsitutes of low-ranked teams. From Table 11, it can be observed that the significant impact on the 'XG' metric by top-ranked teams' substitutes is 0.02, while the low-ranked teams' substitutes have a significant impact of 0.01. Thereby, offensive substitutes of top-ranked teams, have approximately a 100% higher average impact on the metric. However, it must be noted that the average impact of +0.02 expected goals is still not extremely large. Additionally, both by top-ranked and low-ranked teams, no statistically significant impact is generally made on the 'XG' metric in drawing scoreline scenarios.

Table 11: Top/low-ranked teams: Panel data regression where 'XG' is the dependent variable (offensive substituting)

	Coefficients (p-values)			
Variables	**Top_drawing**	**Low_drawing**	**Top_behind**	**Low_behind**
Substitution_dummy	-0.01	4e-03	0.02**	0.01**
Shots_on_target	0.11**	0.15**	0.12**	0.11**
Duels_won	2e-03	2e-04	1e-03	2e-03**
PPDA	-1e-03*	-8e-04*	-4e-04	-1e-03
Invasion	0.48**	0.48**	0.50**	0.40**
Acceleration	-4e-03	-0.01*	6e-03	-3e-03
Crosses	0.01	9e-03*	1e-03	4e-03*
Passes	-1e-03*	-8e-04*	-8e-04*	-1e-03**

Note:
The World & European Championship match data is not incorporated within this analysis for indicated reasons; (*p < 0.05, **p < 0.01)

Table 12: Top/low-ranked teams: Panel data regression (Poisson) where 'Shots on target' is the dependent variable (offensive substituting)

	Coefficients (p-values)			
Variables	**Top_drawing**	**Low_drawing**	**Top_behind**	**Low_behind**
Substitution_dummy	0.10	0.12	0.14	0.12
XG	3.61**	4.06**	2.86**	3.53**
Duels_won	0.07*	0.13**	0.07*	0.06*
PPDA	-0.02	-0.01	-0.02	-0.01
Invasion	2.03*	4.03**	2.74**	3.01**
Acceleration	0.14	-0.71	-0.32	-0.13
Crosses	0.07	0,06	0.18**	0.14**
Passes	5e-03	0.01	0.02*	0.01*

Note:
The World & European Championship match data is not incorporated within this analysis for indicated reasons; (*p < 0.05, **p < 0.01)

The results in Table 12 indicate that both offensive substitutes of top-ranked teams and low-ranked teams seem to have no significant impact on the other offensive metric 'Shots on target', either when drawing or when behind.

Defensive substitutes of top-ranked teams seem to have a significant positive impact on the total amount of duels won, both when drawing and ahead (Table 13). Generally, low-ranked teams' defensive substitutes seem to have a positive significant average impact on the metric, but only in the drawling scoreline scenario. Also, the impact by the defensive substitutes of the low-ranked teams is lower than the impact of the top-ranked teams'. Regarding the top-ranked teams, the average expected increase in the total duels won after defensively substituting is equal to exp(0.11) = 1.12 (12% increase) when drawing, and exp(0.08) = 1.08 (8% increase) when ahead. Regarding the low-ranked teams, when drawing, the metric is expected to increase by exp(0.08) = 1.09 (9% increase) as an effect of defensively substituting.

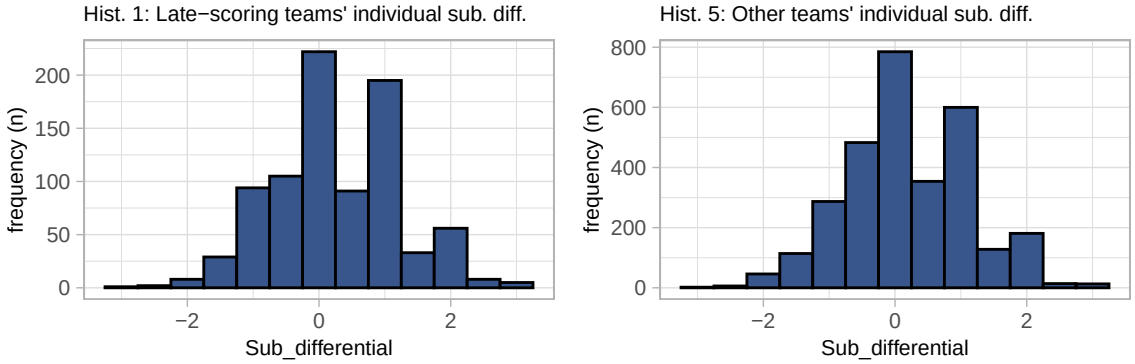
Table 13: Top/low-ranked teams: Panel data regression (Poisson) where 'Duels won' is the dependent variable (defensive substituting)

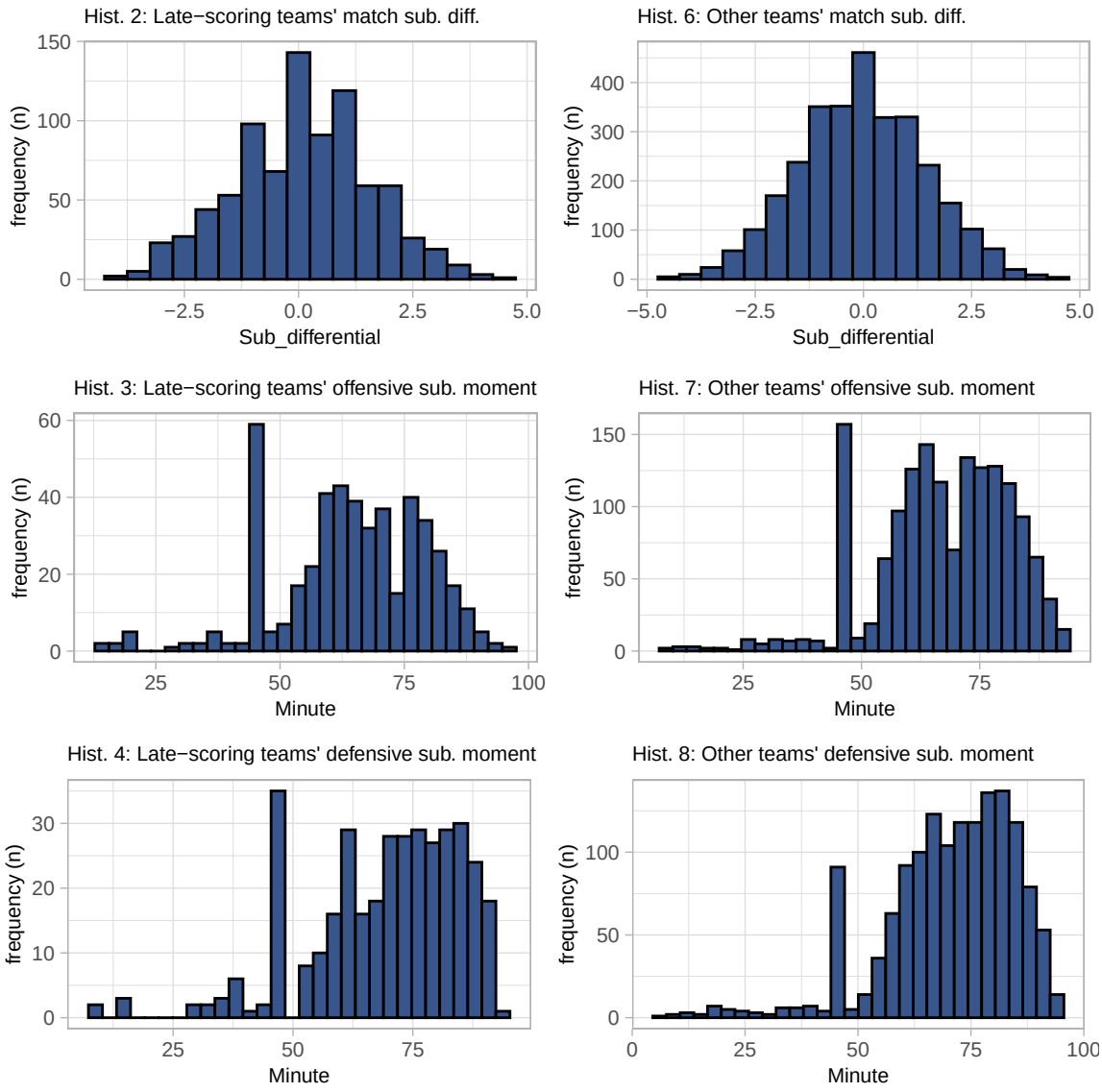
	Coefficients (p-values)			
Variables	**Top_drawing**	**Low_drawing**	**Top_ahead**	**Low_ahead**
Substitution_dummy	0.11*	0.08*	0.08**	0.07
Shots_on_target	0.01	0.10*	0.07**	0.03
XG	0.21	0.12	-0.01	-0.06
PPDA	-0.01**	-0.01**	-0.01**	-0.01**
Invasion	-1.73**	1.02**	-0.11	-0.15
Acceleration	0.02	0.12**	0.05**	0.05
Crosses	0.12**	0.07**	0.10**	0.16**
Passes	2e-03	0.01**	4e-03**	7e-03**

Note:
The World & European Championship match data is not incorporated within this analysis for indicated reasons; (*p $<$ 0.05, **p $<$ 0.01)

6.4 Comparison of substitution strategies of late-scoring teams versus other teams'

To begin with, several histograms are illustrated which contain information about the distributions of the statistics of interest. Note that two groups will be compared with each other: the late-scoring teams, and all other teams. To answer the **sixth research question**, whether a statistically significant difference exists between the groups' substitute strategies, the Mann-Whitney U test will be performed on the groups' data on each statistic. Thereby, it can be derived if some strategy is potentially related to the late-scored goals.





As can be observed from the histograms, it is difficult to visually detect strong differences between the groups wherefrom immediate conclusions can be drawn. Therefore, the Mann-Whitney U test will be performed to bring more clarity to this issue.

To start with, the means of both groups on each statistic are computed, which will provide some initial insights. The means of 'Hist. 1' and 'Hist. 5' are 0.26 and 0.20, respectively. This implies that late-scoring teams, on average, individually substitute slightly more offensively, independent of how the opposing team substituted (the opposing team is incorporated in the 'match substitution differential'). The means of 'Hist. 2' and 'Hist. 6' are 0.09 and -0.03, respectively. This implies that the late-scoring teams, on average, substitute slightly more offensively than their opposing teams. The means of 'Hist. 3' and 'Hist. 7' are 64.4 and 67.2, respectively. This implies that

late-scoring teams, on average, utilize their offensive substitutions earlier than the other teams. And the means of 'Hist. 4' and 'Hist. 8' are 69.0 and 70.5, respectively. This implies that late-scoring teams, on average, utilize their defensive substitutions earlier than the other team. Thus, differences in substitution strategies of late-scoring teams and other teams seem to occur. The next stage is to check if these differences can be regarded as statistically significant by performing the Mann-Whitney U test.

The Mann-Whitney U test shows a p-value of 0.075 (>0.05) when performed on the difference of the groups' values containing information about the individual substitution differential (Hist. 1 and 5). Thereby, it can be concluded that there exists no statistically significant difference on a 5% significance level. However, if a 10% significance level is used, it would be recognized as statistically significant.

The Mann-Whitney U test shows a p-value of 0.02 (<0.05) when performed on the difference of the groups' values containing information about the match substitution differential (Hist. 2 and 6). Thereby, it can be concluded that there exists a statistically significant difference on a 5% significance level.

The Mann-Whitney U test shows a p-value of 0.0001 (<0.05) when performed on the difference of the groups' values that describe the timing of the offensive substitutions of the two groups (Hist. 3 and 7), implying there exists a statistically significant difference between the two.

Regarding the other comparison between the groups' timing of the defensive substitutions (Hist. 4 and 8), no significant differences are found by the statistical test, as the corresponding p-value equals 0.22.

7. Conclusions

This section contains a summary of the performed research, which will provide answers to the initial research questions followed by a discussion.

7.1 Summary

Within this research, the Random forest method appeared to be a well-predicting and interpretable classification model. Herefore, the method was preferred over other machine learning methods for

the task of deriving the variables' importance measures within the classification problem. The test set accuracy was approximately 57%. The latter shows that it is difficult to correctly predict a football match outcome based on the available match statistics variables. This could be explained by the nature of the very dynamic game of football, in which many factors exist which influence the match outcome, where everyday luck is also one of them. The variable importance measured by the Random forest method provided the most important variables in predicting the match outcome in the categories 'Lose', 'Draw' and 'Win'. Offensively, the 'XG' and 'Shots on target' metrics seemed to be most important. Defensively, 'Duels won' and 'PPDA' seemed to be most important. Nonetheless, the 'PPDA' metric seemed to be especially important within the prediction task of draws. Interestingly enough, the 'Invasion index' and 'Acceleration index', both advocated by **Pappalardo et al. (2019)**, did not seem to be extremely decisive in the classification task. Neither was the 'Passing indicator' variable, which is a metric that was also proposed by **Pappalardo et al. (2015)**.

The second and third research questions refer to the general impact of either offensive or defensive substitutions. In general, the scoreline at the moment of substituting is related to the impact of the respective substitutions. Since the metric 'XG' is found to be a very important offensive performance indicator, it is interesting to note that the coach can make a positive impact on the metric by substituting offensively. However, the impact is still small as can be observed in Table 8. Regarding the 'Shots on target' metric, in cases of drawing the average impact seems to be bigger than when behind.

The fourth research question refers to the magnitude of the impact on the key performance indicator variables. From previous results, it can be concluded that coaches, on average, have a significant impact by implementing their strategy adjustments dependent on the scoreline of the moment of substituting. However, their impact on the found most important performance indicator 'XG', is limited on average. Especially, the impact on the total shots on target and duels won was found to be quite significant (ranging from a 8-21% increase). In summary, it can be stated that coaches can make a statistically significant impact on the key performance indicator variables by utilizing their substitutions. Thereby, the coaches' substitution strategy influences the subsequent match outcome.

The fifth research question on hand is about analyzing the difference between top-ranked and low-ranked teams' substitutions impact. By doing so, it can be concluded that, on average, the coaches of

top-ranked teams make a greater impact by their substituting strategies than coaches of low-ranked teams. The finding that top-ranked teams have approximately a 100% higher average impact on 'XG' confirms the previously-made prejudice that substitutes of top-ranked teams generally have a greater impact. Most likely, this impact difference can be explained by the fact that top-ranked teams have a wider and better player selection. Namely, at top-ranked teams, more highly-skilled players are positioned on the bench during a game. These highly-skilled players seem to be able to make a bigger impact, whilst the substitutes of the low-ranked teams also seem to have some positive significant impact on the 'XG' and 'Duels won' metrics. Note again that **Mohr et al. (2003)** found that top-class players generally perform more high-intensity running periods during a game than players at a less elite standard, which may also partly contribute to the distinction. Another possible explanation for the distinction is that coaches of top-ranked teams are better strategists, whereby they have a greater ability to make decisive strategy adjustments throughout the match. By the obtained research results, it can be concluded that the top-ranked teams' budget allocation to its wide player selections may be well worth its money since more in-game impact is made by the top-class substitute players. Note again that the made impact is still relatively low. Hence, the consideration of whether the money is worth the gained advantage is a decision to make by the club management.

Finally, the last research question about the difference in substitution strategies of late-scoring teams versus all other teams can be answered. Based on the results in Section 6.4, late-scoring teams are utilizing offensive substitutions structurally earlier in the game than the other teams. This is in line with the research done by **Myers (2012)**, who claims that coaches should utilize their substitutes in early stages of the game. Regarding defensive substitutions, there seems to be no significant difference between the two groups. There also seems to be a statistically significant difference (at a 5% significance level) in the match substitute differential of the two groups. This means that, when comparing within-match differences (substitution differential between two opposing teams), it can be stated that late-scoring teams generally substitute *more offensively* than their opponents, which is an interesting finding. Nevertheless, on a 5% significance level, on average, late-scoring teams do not seem to differ significantly in terms of their individual substitution differential. However, if a 10% significance level is used, they do. Thus, in summary, the late-scoring teams generally substitute more offensively and earlier than the other teams. This finding also supports the earlier finding that substitutes can make an impact on the match outcome. The latter is in line with the

finding of the substitutes' positive, yet small, impact on the 'XG' metric. Thus, generally speaking, the results show that substitutes make an impact on the metric, but the impact is conditional on the current scoreline. Therefore, it can be stated that coaches can make a difference by their substitution strategies, where coaches of top-ranked teams make generally a greater impact than coaches of low-ranked teams. In all cases, the scoreline at the moment of the strategy adjustment is related to the subsequent impact.

As a conclusion, the claims made by **Schacter (1999)** and **Silva et al. (2016)** can now be reviewed. Hence, by this research, can it be confirmed that people's memory and confirmation bias play a role within the perception of the substitutes' impact such that only outstanding events are remembered to solidify their previously held opinions? This psychological theory generally does not seem to hold by the results of this research since in several scoreline scenarios substitutes are able to structurally make a difference. Hence, when football fans tend to remember several outstanding match events where the substitute turned out to be the team's hero, by this thesis' model it can be statistically backed-up if they are rightly in doing so.

7.2 Discussion

As a side note, it must be noted that the variable 'substitution dummy' is never able to fully describe the real substitutes' impact. Partly, this is because some aspects of the game cannot be translated into data. Furthermore, it is unlikely that only the substitute will be the cause for all events to happen after the moment he is substituted. Many events are possible to happen throughout the match, whilst the substitution being only one part of it. Thereby, it is unrealistic to assume that all fluctuations in the match statistic variables are always fully caused by the substitute. In reality, many factors can have an influence on those variables over the course of a match, where everyday luck is also one of them.

Another aspect to keep in mind is that when the end of the match approaches, losing teams are generally willing to take more risks in order to retrieve their losing situation. At the same time, usually, substitutions have been made, which will result in the 'substitution dummy' variable being equal to 1. When the 'XG' metric increases because of more successful risks are taken by a team, the model will assign this credit to the substitutes, whilst this may be partly wrong. The same holds for winning teams who want to retain their winning situation. In such a scenario, when the end of the match approaches, the winning team might be tempted to "lean backwards" and focus

primarily on defending to secure the result. In such a scenario, the total amount of duels won may be increased due to their different playing style, whereas usually also defensive substitutions have been made. Thereby, the model will again assign this credit fully to the substitutes, whilst this may be partly wrong.

As a final conclusion, it can be stated that it remains difficult to fully chart the substitutes' impact. However, within this research, a great attempt has been made. The useful managerial insights which follow from the obtained result can be relevant either for the general management of clubs or for the club's coaches. Thereby, this research is relevant within a sports context. Nevertheless, a similar research could be paralleled within a business context where human resource management practices are analyzed, or even within a medical context where the productivity of successive hospital personnel shifts are analyzed.

References

Bartling, B., Brandes, L. & Schunk, D. (2015). Expectations as Reference Points: Field Evidence from Professional Soccer. *Management Science*, 61(11):2646-2661. https://doi.org/10.1287/mnsc.2014.2048

Bornn, L., Cervone D, & Fernandez J. (2018). Soccer analytics: Unravelling the complexity of "the beautiful game". *Significance*, 15(3):26-29. doi: 10.1111/j.1740-9713.2018.01146.x

Bradley P.S. & Noakes T.D. (2013). Match running performance fluctuations in elite soccer: indicative of fatigue, pacing or situational influences? *J Sports Sci*, 2013;31(15):1627–38.

Bradley P.S., Lago-Peñas, C. & Rey, E. (2014). Evaluation of the match performances of substitution players in elite soccer. *Int J Sports Physiol Perform*, 2014;9(3):415–24.

Del Corral, J., Prieto-Rodriguez, J. & Pestana Barros, C. (2008). The Determinants of Soccer Player Substitutions: A Survival Analysis of the Spanish Soccer League. *Journal of Sport Economics*, 9(2):160-172. doi: 10.1177/1527002507308309

Emanuel E.J., Persad G., Upshur R., et al. (2020). Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine.*, 382(21):2049-2055. doi:10.1056/NEJMsb2005114

Gomez, M., Lago-Penas, C. & Owen, L.A. (2017). The influence of substitutions on elite soccer teams' performance. *Journal International Journal of Performance Analysis in Sport*, 553-568.

Hills, S.P., Barwood, M.J., Radcliffe, J.N., Cooke, C.B., Kilduff, L.P., Cook, C.J. & Russell, M (2018). Profiling the Responses of Soccer Substitutes: A Review of Current Literature. *Sports Med* 48, 2255–2269 (2018). doi: https://doi-org.eur.idm.oclc.org/10.1007/s40279-018-0962-9

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: with applications in R. *New York: Springer.*

Link, D. (2018). Data Analytics in Professional Soccer. *Wiesbaden: Springer Vieweg.* doi: https://doi-org.eur.idm.oclc.org/10.1007/978-3-658-21177-6

Mohr, M., Krustrup, P. & Bangsbo, J. (2003). Match performance of high-standard soccer players

with special reference to development of fatigue. *J Sports Sci*, 2003;21:519–528.

Myers, B.R. (2012). A proposed decision rule for the timing of soccer substitutions. *Journal of Quantitative Analysis in Sports*, 8, Article 9.

Padron-Cabo, A., Rey, E., Vidal, B., & García-Nuñez, J. (2018). Work-rate Analysis of Substitute Players in Professional Soccer: Analysis of Seasonal Variations. *J Hum Kinet*, 65: 165–174. doi: 10.2478/hukin-2018-0025

Pappalardo, L, Cintia, P., Pedreschi, D. & Giannotti, F. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. *Paris: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA'15)*. doi: 10.1109/DSAA.2015.7344823

Pappalardo, L., Cintia, P., Rossi, A. et al. (2019). A public data set of spatio-temporal match events in soccer competitions. *Sci Data* 6, 236. https://doi-org.eur.idm.oclc.org/10.1038/s41597-019-0247-7

Rey, E., Lago-Ballesteros, J. & Padrón-Cabo A. (2015). Timing and tactical analysis of player substitutions in the UEFA Champions League. *Int J Perform Anal Sport*, 2015;15:840–850.

Schacter, D.L. (1999). The seven sins of memory: insights from psychology and cognitive neuroscience. *American Psychologist*, 54, 182-203.

Silva R.M. & Schwarz T. B. (2016). Analysis of Substitution Times in Soccer. *Journal of Quantitative Analysis in Sports, De Gruyter*, vol. 12(3), pages 123-124, September.

Trainor, C. (2014). Smart Use of Substitutes Can Make A Difference. *Statsbomb*.

Wu, P., Fang Y., Guan Z., et al (2009). The psychological impact of the SARS epidemic on hospital employees in China: exposure, risk perception, and altruistic acceptance of risk. *Can J Psychiatry*, 54: 302–11.

Appendix

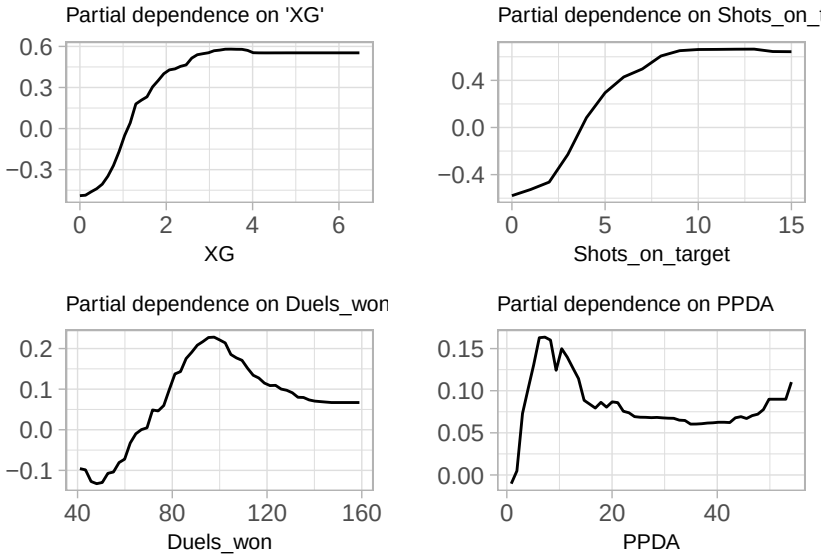


Figure 9: Partial dependence plots of most important variables on match outcome class 'Win'

Table 14: Confusion matrix by OLR

		Reference		
Conf.Matrix	Class	Lose	Draw	Win
Prediction	Lose	219	98	81
	Draw	0	0	0
	Win	73	88	211

Table 15: Confusion matrix by Random forest

		Reference		
Conf.Matrix	Class	Lose	Draw	Win
Prediction	Lose	216	100	75
	Draw	15	15	12
	Win	61	71	205

Table 16: Confusion matrix by SVM (radial kernel)

		Reference		
Conf.Matrix	Class	Lose	Draw	Win
Prediction	Lose	224	99	77
	Draw	7	9	5
	Win	61	78	210