2020

The characteristics of winning football matches: Determining a model that can predict the match outcome of football matches

B.A.G.M. Steijns
Student number: U1267555

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. G. Spigler
dr A. Alishahi

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2020

Preface

It was a pleasure for me to write this thesis since I was able to combine my passion for football with the knowledge I have gained at Tilburg University on data analytics. I would like to express my most sincere gratitude to my supervisor professor Giacomo Spigler for his valuable feedback and interactions during this research. Lastly, I would like to thank the University staff, second reader, my family & my girlfriend for the support during my six years of studies at Tilburg university

Abstract

This research studied the effect of several match-related, team-specific and player-specific features on the outcome of a football match. The aim was to predict the FTR (Full-time result) variable by building three models based on various publicly available statistics. The football leagues that were studied in this research are the Top five football leagues in the world: English Premier League, Spanish Liga BBVA, German Bundesliga, Italy Serie A & French Ligue 1. The first goal of this study is to discover new features that have a high correlation with the match outcome. Secondly, using these features we try to improve upon the prediction results of previous researches. The classifiers that have been used in this study are the Random Forest, Gradient boost and Logistic regression. The gradient boost classifier outperformed the other two classifiers in most of the predictions. Match statistics, player statistics and team statistics were obtained for the seasons 2008/09 until 2015/16. Feature selection was performed using the Pearson correlation coefficient, while applying feature extraction to the dataset led to newly engineered featured.

The top three features that yielded the best correlations with the FTR variable were: Difference in League points, Difference in League position last season and the Difference in overall rating between players in the home- and away team. Our model predictor based on solely match-statistics using gradient boost obtained the highest accuracy for the Spanish league (0.56), while the player-statistics model also obtained decent accuracies. The predictor based on player statistics using logistic regression performed the best for the English premier league with an accuracy of 0.54. The model based on team-specific attributes was not very successful in predicting the match outcome. Merging the match-based features and player-based features did not lead to a significant increase in accuracy. Except for the leagues that were least predictable in the sub-questions, like the German and French leagues, higher accuracy was obtained in this model.

Table of Contents

1. Introduction

In today's society, the game of football has become a lot more than just a game. The importance of football in the world is huge and as the popularity of the sport is increasing, the revenues that are associated with winning football matches also increase. One of the football leagues that will be studied in this research is the English Premier League. The football club that will win this league will generate huge revenues. Last season, Liverpool FC were crowned champions and generated a record-breaking 175 million pounds for playing in the premier league this season (Moore, 2020). To put this figure in perspective, the Premier League winners of 10 years ago, Manchester United, generated total revenues of 52 million pounds in 2010 (Harris, 2010). So, this is an increase of 236% in terms of revenue generation in an European football competition. Next to the increase in possible revenues for football clubs, there has also been an increase in costs over the years. One of the biggest costs for football clubs is acquiring players. 10 years ago, the most expensive transfer was Cristiano Ronaldo who transferred from Manchester United to Real Madrid for 80 million pounds (Transfer market, 2020). Today the record stands at 198 million pounds for Neymar who was bought by Paris Saint Germain from Barcelona. This increase in revenues and costs over time reveals the growing importance of football.

It is imaginable that in such an important sport you want to leave as little to chance as possible. However, over the years, the sport of Football has been quite unpredictable. The unpredictability in football is mainly caused by the great amount of variables that can influence a match outcome. The most important variables are the twenty-two players that are on the field. However, it is also vital where the manager places them on the field. Next to this, certain factors that take place a bit more in the background, as the Training facility, location of the match, and the weather can also have an influence on the outcome of the match. This unpredictability in the game of football combined with the increasing interest in the game displays the need for research in football, especially the need for data-driven research.

The rise of big data and machine learning in this century has opened up possibilities for data analysis in fields where this was unthinkable in the past. Currently, data analytics is extremely popular in football as the rise in data and machine learning techniques have opened up all kinds of possibilities for value creation in the sport. In 2012 the Football club Arsenal bought the data analytical company called StatDNA for around $4 million dollars (NY Times, 2017). They purchased this company with the aim of trying to get an advantage over their competitors by using data analytics. These data analytics could help the club make decisions on the selection of the team. The growing demand for these data analytics in football can be seen by the rise in data companies like Scisports and Opta. They are being hired by football clubs to collect data and use this data to generate valuable insights for the club.

This research attempts to provide an in-depth analysis of various features which can have a direct influence on the match outcome. After performing this analysis, the research question that needs to be answered is:

"To what extent can investing in high-depth data analysis be beneficial for the prediction of match results?".

There have been a lot of researches done in this area in the past decades, some of these researches are highlighted in the next section. However, the rise of big data in the last couple of years has opened up new opportunities for research on this topic. Players are nowadays being tracked via GPS devices, and their bodies are being monitored during football matches with heart rate sensors. These are just a couple of examples of data extraction that will present opportunities for new valuable insights to be generated through research. Besides the increase in data availability, the ever-changing game of football also requires up to date analysis since the way football is being played today differs a lot from 20 years ago. For example, the game has become less physical and the pace of the game is a lot higher today. These changes in play increase the need for up to date research on this topic. The most important conclusions drawn from previous researches will be discussed in the related work section. Most researchers have only had the ability to use limited data in the form of previous match results and league rankings. With the dataset acquired in this research, there will be a huge amount of possibilities in terms of input features for the prediction, as will be described in the dataset section.

This research aims to provide valuable insights into the field of data analytics in football. These valuable insights have to established using feature presentation, feature selection & feature extraction. After which the newly created dataset will be tested against a benchmark model to see the outcome. These techniques will prepare the data to see if in-depth match factors, as in player-specific attributes, team-specific attributes and detailed match events have an influence on the match outcome. This will be done by using data visualization and eventually model prediction. The results of this study are in line with previous research done on predicting the outcome of football matches. As predicted by previous researches, match statistics have the highest correlation with the match outcome, while player statistics also prove to correlate significantly with the match outcome. The team-specific attributes that are studied have a significant correlation with the match outcome, however, they do not improve the predictor a lot. This is mainly due to the fact that they overestimate the chances of a home win. This research will make a contribution to the existing literature by studying new statistics and engineering new features. The newly created features that have a highly significant correlation with the match outcome could be studied further in other researches to improve the prediction of the match outcome.

The thesis is displayed in the following manner. Section II provides an overview of the main research question. In section III, the related work in the research field of predicting match outcomes is described. Section IV displays the experimental set-up and the methods used in this research. The results of the various predictors are presented in section V, which is split into separate subsections answering each sub-question. Lastly, sections VI and VII display the discussion and conclusion following the results obtained from this research, whereafter Section IIX suggests possibilities for future research.

2. Main research question and hypotheses

This research aims to see if the outcome of a football match can be predicted by building a model that can accurately predict the match result. This will be done via a careful feature selection and extraction process, after which the significant features will be used in various classifiers to see which model works best for our dataset, therefore the main research questions will be:

RQ: *"To what extent can investing in high-depth data analysis be beneficial for the prediction of match results?".*

To examine the research question, Three sub-questions have to be answered. Firstly, the relationship between historical data on previous matches and the match outcome will be studied. As previously stated in the literature, it has been proven that historical match data has a positive influence on predicting the match outcome. In this research this will be studied by creating new features from the previous match results of each team. these features will be checked for correlation and the relevant features will be used in the final predictor of this model. Sub-question one is displayed below.

SQ1*: Do historical match related events contribute to the prediction of the outcome of a football match ?*

After reviewing historical data, the second and third sub-question will use actual in-season data. The second sub-question views specific team-statistics data that are based on data from the game EA Sports FIFA. This data will be used to view if a certain way of playing for a team can help the model in predicting the outcome of a match result. At first, the statistics from EA Sports FIFA will be checked for significance after which the second sub-question will be studied.

SQ2*: Do Team-specific attributes contribute to the prediction of the outcome of a football match ?*

The third sub-question examines various player-specific statistics in order to see if they have a direct correlation with the match outcome Similar to the team-statistics, this data is gathered via EA Sports FIFA and controlled for significance before being used in the final predictor.

SQ3*: Do Player-specific attributes contribute to the prediction of the outcome of a football match ?*

3. Related work

This chapter provides related work on this topic and can present various insights in which feature selection and model prediction methods that can be applied to this research. Lots of research has already been performed in the field of data science in Football. Several pieces of research that are relevant to this paper will be listed below. The overall conclusion in most of the previous literature is that the prediction models perform quite well compared to the baseline. However, it remains difficult to define a model that can accurately predict the match outcome. The following sections will present overviews on previous studies concerning the three different research questions.

The first research question focusses on the match statistics. Match statistics have been the core dependent variables of most previous researches that try to predict the outcome of a football match. One of the main reasons that there is so much interest in research related to football statistics are the gambling possibilities. According to data from Research and Markets (2020), the sports betting market was valued at US $85.0 billion in 2019. Sports betting is the activity in which someone can put a wager on the outcome of a sports match. In order to correctly predict these matches, bettors want to leave as little to chance as possible. This is why the goal of multiple researches was to beat the bookmakers. The study of Gomes et al. (2015) is one of these studies that have made an attempt to make profits from betting on sports. They designed a decision support system that achieved a 20% return when tested on the English premier league. This research collected various data on twelve seasons of English Premier league football. Their data consisted of in-match statistics like shots, fouls, cards, goals. In order for this data to be workable, new features were created by averaging the in-match statistics. Gomes et al. (2015) used three learning algorithms to predict the match outcome: SVM, decision tree and Naïve bayes. After successfully building their model, they could predict the match outcome with an accuracy of 51%.

Another study relevant to this research was done by Constantinou et al. (2012). Similar to the study of Gomes et al. (2015), they tried to predict the outcome of football matches played in the English Premier League. A Bayesian network was used in which the subjective features represent the important factors for prediction which historical data failed to capture. There are four main features that they have created to predict the match outcome; strength, form, psychology and fatigue. The first feature is based upon historical data as in the league ranking of the previous year and the outcome of previous matches. Features 2,3 and 4 mainly rely on subjective information. This subjective information can revise the outcome measured via feature 1. Constantinou et al. (2015) are one of the first studies in this area that uses subjective information in their predictor. This turned out to be very successful. Similar to Gomes et al (2015), they were also able to beat the odds presented by bookmakers. From this research, it can be concluded that subjective information can help the prediction model. However, in our research, we do not have this data available so it has to be viewed if a fully objective model can also obtain similar accuracies.

Similar to previous studies discussed, Ulmer et al. (2013) also used football match data to predict the outcome of football matches in the English Premier League. The features that they used in their research are the ranking, form, goal difference, and home advantage of the team. The Random Forest classifier, which will also be used in this research, and SVM classifier were used to test the model, they obtained accuracy scores of 0.50 for Random forest and 0.51 for SVM. Ulmer et al. (2013) concluded that in order to prevent overfitting of the model, the complexity of the model should be reduced.

One of the most recent researches on this topic was performed by Baboota & Kaur (2018). Their aim was to beat the bookmakers by trying to accurately predict the English Premier League. The pre-processing using pandas in python and modelling using sckit-learn in this research corresponds to the pre-processing and modelling that will be done in this research. Moreover, they used feature engineering and exploratory data analysis to create a feature set that they used for their predictions. The results of their model were quite promising, especially using the gradient boost. However, they could not outperform the bookies. For further research, it is recommended that the quantity of the dataset should be larger as this will probably result in more accurate analysis.

The research of Tax et al. (2015) studied a different league. They predicted the Dutch football competition using a machine learning approach. Similar to the studies described above, match statistics were used in their research to obtain an accurate predictor. Tax et al. (2015)

built their own dataset with a great number of candidate features. The features used varied from goals scored, previous match outcomes, team form, earlier encounters versus similar team, etc. After collecting all the data, 45 features were made that could help predict the outcome of the match. When creating a great number of features based upon match statistics solely, the problem of multicollinearity can arise. This is a problem that can also occur in our research as there will be lots of dependent variables that can have a high correlation. In order to address this problem, Tax et al. (2015) used various dimensionality reduction techniques. The best results were obtained using principal component analysis. This is a dimensionality reduction technique that will also be used in our research since we will use similar data as Tax et al. (2015). PCA is a multivariate analysis method that is used to reduce the dimensionality of datasets and increase interpretability while minimizing the loss of information. Next to this, The player-specific features are also highly correlated with each other so a dimensionality reduction technique similar to the one used by Tax et al. (2015) will be useful for answering the third research question.

This research will view if the regular prediction model based on match statistics can be improved by using other statistics, like team and player attributes. Multiple other pieces of research in the past also incorporated other statistics besides the match statistics. One of these researches is the study of Kampakis et al. (2015). Their main goal was to beat the odds provided by bookmakers on cricket matches They had lots of data available, varying from match statistics to team and player specific attributes, which led to an extensive feature selection process. The feature selection methods that were used are Chi-square, mutual information and Pearson correlation scores. Afterwards, they viewed if PCA (Principal Component Analysis) prior to the learning would be beneficial. To verify which input features are relevant, the Pearson correlation scores seemed most efficient. Whilst for the learning task, naïve Bayes produced the best outcomes. The conclusions that can be drawn from this research are that the pre-processing and feature selection process are really important steps in predicting the outcome of sports matches.

The research of Adam (2016), uses player-specific statistics in a generalised linear model as the most important feature to predict the outcome of an international football match. They participated in a prediction challenge in the sports Analytic Lab of the KU Leuven with two main tasks. The first task consisted of predicting the match outcome of a football match.

For the second task, they had to predict which team would advance in the tournament. They obtained some good results, however, they suggested that for future research it would be good to add additional features to the dataset like passes and assists. Further recommendations are that due to the multiplicity of features, some feature selection process is necessary, which is also applicable to the player statistics in this research. Tavakol et al. (2016) also tried to predict the match outcomes of the Euro 2016 by using a linear model. In their research they used quite a variety of features which makes this research interesting; Features that they used to include the Countries ranking and Player statistics varying from market value, age, career matches to goals. However, their data was quite limited and therefore their model did not have the best outcomes. Their expectation is that when there is enough historical data available, their "simple" approach can built a good predictor for match outcomes.

The last study that will be discussed is the study of Min et al. (2008). In this research, a model was made that can predict the outcome of football matches by using Bayesian inference and rule-based reasoning together with an in-game time series approach. The interesting part of their research is that they consider a skill variable that is linkable with the formation of each team. They build four Bayesian networks that have to help decide a team their strategy. Lots of general variables are available to them like the average rating of the defence, average rating of midfield and an average rating of the offence. Next to this, more specific variables like finishing stats, chances, man-marking, teamwork, fatigue etc. are also available. All these variables are combined into the following four Bayesian networks; Offensive grade, defensive grade, possession grade and Fatigue. As said before, with these four networks they attempt to discover a team's strategy which can help them predict the match outcome. The similarities between this study and the research of Min et al. (2008) is that we also have a huge amount of skill-based data available to help to predict the match outcome. In our research, we will combine various correlated and significant skill ratings for each position on the field to obtain three final components per position which should benefit the predictor. Please view the appendix for a brief summary of the most important researches mentioned.

4. Experimental Setup and Methods

This section will at first present a description of the dataset and programming languages used Next to this, the theorical framework, methods used and evaluation method are discussed.

Dataset Description

The dataset that is being used in this research is in terms of quantity and quality the best publicly available dataset in this area of research. It has a variety of information on football matches & football players in different leagues. The main reason for choosing this dataset is the large amount of data that is gathered on Football teams, football matches and football players. This will enable this research to hopefully generate new insights in this field of science where previous researches did not have this detailed information available. The dataset is downloaded from Kaggle.com (European soccer database). The dataset is in SQL format and it will have to be converted to python format for the pre-processing part. It contains information of over 25,000 Matches from 11 different European leagues. In this study, the top five leagues (English Premier League, Germany Bundesliga, Italy Serie A, Spain Liga BBVA & France Ligue 1) will be studied as they contain the most information on matches. Next to the in-game statistics tables, the dataset also contains two tables with the team and player statistics per season. These statistics are gathered via the popular game EA Sports FIFA, which has a broad network of data reviewers that closely monitor these statistics. In order to create some extra variables involving match statistics, we needed to create one extra dataset containing the league rankings of all the leagues over the studied period (2008-2015). This data was obtained via Transfermarkt.com (Transfer market, 2020). The league rankings for the corresponding seasons can be found in the appendix.

Software

The programming language of python is used in this research in order to answer the research question. The program that was used to code in python with is Anaconda Navigator, this is an open-source distribution of different programming languages. The main packages that were used in this research are NumPy for the pre-processing and Sci-kit learn for the prediction and feature selection process. In order to retrieve the match statistics that are presented in XML format, the xml.etree.ElementTree package had to be imported. The classifiers used to predict the match outcome were Logistic Regression and Random Forest using sklearn and gradient boost using xgboost. An overview of all the used packages can be found in the appendix.

Theoretical Framework

The theoretical framework that is displayed below in Figure 1 provides for each sub-question an example of how the model is built and the experiment is executed. The three different datasets are studied, a feature selection process will decide the variables that we will use in the final predictor. These variables will be transformed into new features using feature extraction and PCA. This feature set will be split into a training and test set (80/20). This model will be tested on three classifiers to see which classifier performs best.

Figure 1: Theorical framework

This figure presents a roadmap of how this research will be conducted. At first, the data is gathered. A good feature selection and extraction process will generate new features that will predict the match outcome using one of the three classifiers.

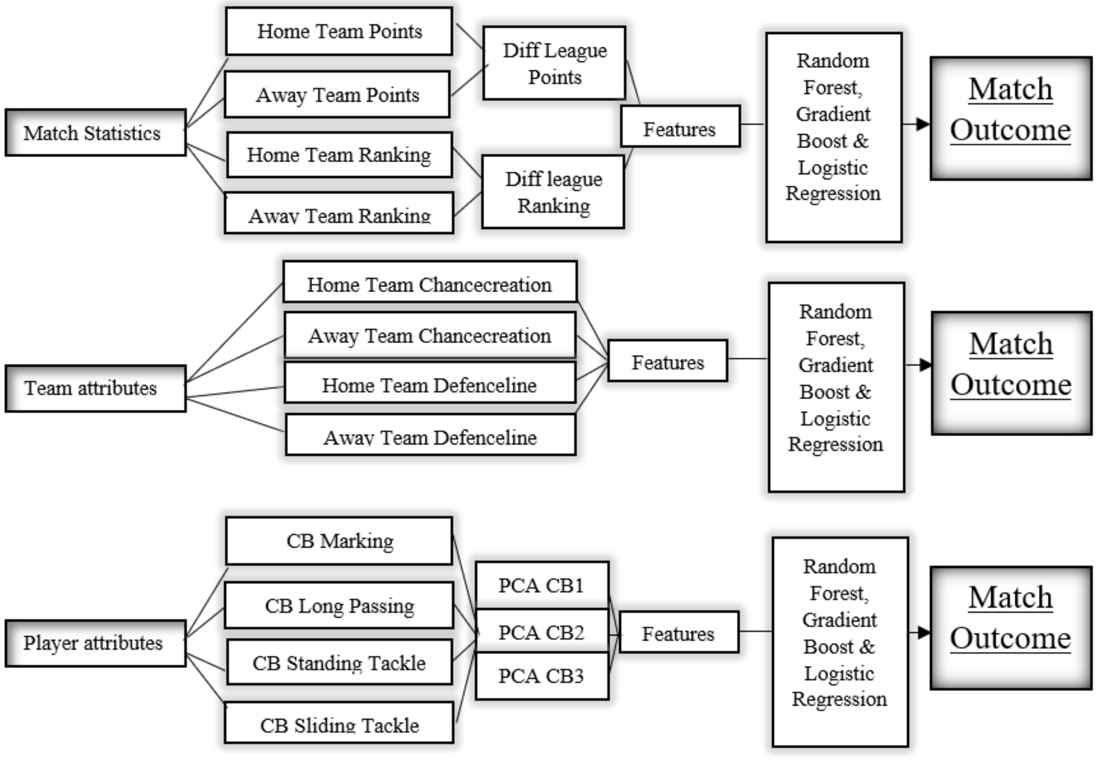


Figure 1. Expected model of the effect of Match statistics, team attributes and player statistics on the match outcome of a football match.

Pre-processing

The following three sections will briefly explain the work that had to be done before the prediction model. The sections that will be discussed are the Data Cleaning, Feature Selection

& Feature extraction. Each section will explain the necessary steps taken to investigate the three sub-questions.

Data cleaning

In total there were 7 tables in the dataset that we retrieved from Kaggle and 1 table gathered via Transfermarkt. From these seven tables, three tables were used to answer the first sub-question. In order to retrieve valuable insights from the SQL dataset we first had to extract the match statistics table from the dataset. This table contains information on the outcome, the week, the season and league of all matches in the dataset. Furthermore, this table contains data on various in-match statistics. These statistics are in XML format so they needed to be translated to real values and columns in a pandas data frame. After this translation has been done, the final dataset was split per league. This was done because the research also wants to see if some leagues are more predictable than other leagues.

Table 1: The description of the original match statistics in the dataset

Data Cleaning: Match Statistics		
Variables	**Description**	**Type**
Match_api_id	Id of the Match	Integer
Date	Date the match was played	Datetime
HomeTeam	Name of the home team	Object
AwayTeam	Name of the Away team	Object
FTHG	Home team goal	Integer
FTAG	Away team goal	Integer
FTHS	Home team shot on target	Integer
FTAS	Away team shot on target	Integer
HOS	Home team shot off target	Integer
AOS	Away team shot off target	Integer
HYC	Home team yellow card	Integer
AYC	Away team yellow card	Integer
HRC	Home team Red card	Integer
ARC	Away team Red card	Integer
HCR	Home team crosses	Integer
ACR	Away team crosses	Integer
HCO	Home team corners	Integer
ACO	Away team corners	Integer
HPO	Home team possession	Float
APO	Away team possession	Float

The last step of the data cleaning was splitting the data frame per season. The need for this step will be seen in the feature extraction part where new features are created.

Before the in-match statistics could be used in our code, some alterations needed to be made. The first trial was done on the English premier league and no data was missing there. However, in other leagues, the data was missing quite some matchdays per season which made the code not work properly. This was solved by either removing the season when more than half of the match weeks were missing or by adapting the code (NumPy arrays and for loops) so smaller seasons (seasons with full match weeks missing) could also be measured. After cleaning the data, the features seen in Table 1, will be studied to see if they can contribute to this research.

The team statistics are extracted from the Team Attributes dataset. This dataset contains information on team-specific attributes as Pressing style, Chance creation quality, etc. These team stats were already in the correct formatting so they did not need to be converted. However, there team-specific attributes are not available for every season we have in our dataset. In order to see which seasons match with the season where the matches are registered in, a cross-check needed to be performed with the Match table. After comparing the season, we noticed that for six out of the eight seasons we had team statistics. This led to the exclusion of Season 08/09 and Season 12/13 from the team statistics dataset. The other statistics per team were matched against the match result per season and went through to the feature selection and extraction part. Table 2 above displays all the relevant features that remained after the cleaning of this table.

Table 2: The description of the original team statistics in the dataset

Data Cleaning: Team Statistics		
Variables	**Description**	**Type**
HomeTeam	Name of the HomeTeam	Object
AwayTeam	Name of the Away team	Object
Home_team_api_id	ID of the home team	Integer
Away_team_api_id	ID of the away team	Integer
Match_api_id	ID of the match	Integer
date	Date the match was played	Datetime
Season	The season the match was played in	Category
Various Team Statistics (0-100)	buildUpPlaySpeed, buildUpPlayDribbling buildUpPlayPassing, chanceCreationPassing, chanceCreationCrossing, chanceCreationShooting, defencePressure, defenceAggression, defenceTeamWidth,	Integer

Various Team Statistics (Class)	buildUpPlaySpeedClass, buildUpPlayDribblingClass, buildUpPlayPassingClass, buildUpPlayPositioningClass, chanceCreationPassingClass, chanceCreationCrossingClass, chanceCreationShootingClass, chanceCreationPositioningClass, defencePressureClass, defenceAggressionClass, defenceTeamWidthClass, defenceDefenderLineClass	Category

The third sub-questions measures the effect of player-specific statistics on the match results. To answer this question we have to extract three tables from the dataset; Player Attributes, Match and League. Similar to the previous sub-question, the player attributes table has to be merged against the Match table in order to get the statistics of all the twenty-two player that play in one match combined with the match outcome. In order to gain valuable insights in these statistics, a careful feature extraction process is necessary. This will be explained in the feature extraction section. The variables that were used to create the new features are explained in table 3 below.

Table 3: The description of the original player statistics in the dataset

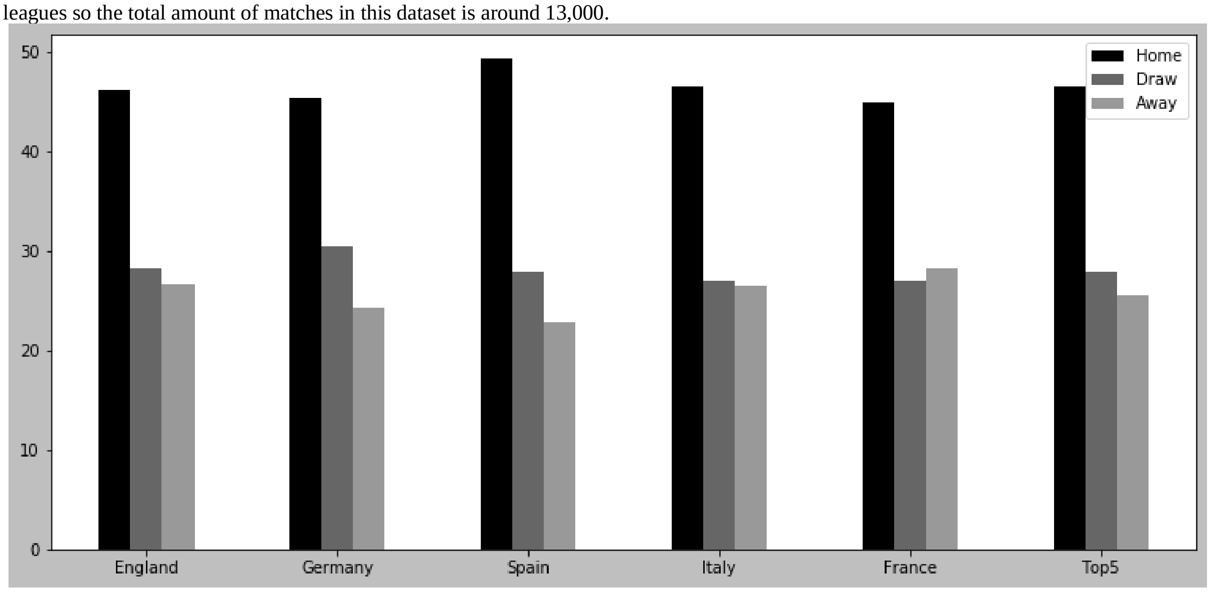
Data Cleaning: Player Statistics		
Variables	**Description**	**Type**
HomeTeam	Name of the HomeTeam	Object
AwayTeam	Name of the Away team	Object
Home_team_api_id	ID of the home team	Integer
Away_team_api_id	ID of the away team	Integer
Match_api_id	ID of the match	Integer
date	Date the match was played	Datetime
Season	The season the match was played in	Category
Home_player_id X1-X11	The ID of the player on each position which can be linked to the FIFA stats	Integer
Away_player_id X1-X11	30 Player-specific FIFA statistics	Integer
Home_player_X1 – X11	For each home player the position in the width of the Field (1-9)	Integer
Away_player_X1 – X11	For each away player the position in the width of the Field (1-9)	Object
Home_player_Y1 – Y11	For each home player the position in the length of the Field (1-11)	Float
Away_player_Y1 – Y11	For each away player the position in the length of the Field (1-11)	Object
30 Player-specific FIFA statistics	Overall Rating, Potential, Crossing, Finishing, Heading accuracy, Short Passing , Dribbling, Volleys, Curve, Free-kick accuracy, Long passing, Ball control, Acceleration, Sprint speed, Agility, Reactions, Balance, Shot power,	Integer

	Jumping, Stamina, Strength, Long Shots, Aggression, Positioning, Vision, Penalties, Marking, Standing tackle, Sliding Tackle	
5 Goalkeeper-specific FIFA statistics	GK Diving, GK Handling, GK Positioning, GK Kicking, GK Reflexes	Integer

After cleaning all the data, the distribution of the target feature can be seen. This distribution is displayed in Figure 2 below. It can be concluded that on average in 45%-50% the home team wins their matches.

Figure 2: Distribution of the target variable FTR

This figure presents an overview of the distribution of the target variable FTR for all the leagues that are being studied. The data consists out of 8 seasons, so around 2000-3000 matches per country are studied are behind these figures. The Top5 league combines the data from all the leagues so the total amount of matches in this dataset is around 13,000.



Feature Selection & Feature Extraction

The most important recommendation from previous researchers on this topic is the importance of a good feature selection and feature extraction process. Lots of statistics on football matches can be a lot more useful when you transform the original features into new features. This section will display for each of the three sub-questions what feature selection and extraction process was performed in this study.

Firstly, the match statistics are studied. These variables require the most feature extraction since these statistics cannot be linked directly with a specific match. The team and player-specific attributes are made at the beginning of the season, so they can then be used for each match in the whole season. The match statistics differ each week following a new match outcome so this variable has to be transformed.

The features that are extracted from the available match data can be seen in Table 4 below. Most of the features in Table 4 are extracted in the same manner. One of these features is the difference in league points, this feature is calculated through subtracting the total points gathered by the away team from the total points of the home team. These total points are re-calculated after each match week. A similar approach is taken for these other features: HTGD, HTONSD, HTOFSD, HTCOD, HTCRD, ATYD, ATRD, HTRD, HTYD, ATCRD, ATCOD, ATONSD, ATOFSD, DiffFormPts. The DiffLP feature is generated by combining the league rankings of each team with their matches in the season after that league ranking. The teams that did not have a league ranking the year before because they were promoted from a lower division, were given a league ranking similar to that of the low-ranked team. All features are also checked for multicollinearity as these statistics can have a high correlation. The shots on target are for example correlated with the total goals since you cannot score a goal without having a shot on target. VIF is a variance inflation factor which can detect multicollinearity. The VIF value display which percentage of the variance is inflated for each coefficient. Features with VIF values over 10 are excluded from the research as this is problematic for the collinearity.

Table 4: The description of the final match statistics in the dataset

Data Cleaning: Match Statistics		
Variables	**Description**	**How**
Match_api_id	Id of the Match	NA
FTR	Full Time Result	Made by defining comparing FTHG with FTAG (H, A, D)
HM1-HM3	Home form in the last 3 weeks	Made by looking at the FTR column for previous matchweeks of each club
AM1-AM3	Away form in the last 3 weeks	Made by looking at the FTR column for previous matchweeks of each club
HTGD	Home Team Goal Difference	HTGS (Goals Scored) – HTGC (Goals Conceded)
ATGD	Away Team Goal Difference	ATGS (Goals Scored) – ATGC (Goals Conceded)
HTP	Home team points in season	Cumulated Home Team points collected so far in the season
ATP	Away team points in season	Cumulated Away Team points collected so far in the season
DIffFormPts	Difference in points in the last matches	HTFormPts (Form of the home team) – ATFormPts (Form of the away team)
DiffLP	Difference in League position last season	HomeTeamLP – AwayTeamLP
DiffPts	Difference in League points this season	HTP (Home Team Points this season) – ATP (Away Team Points this season)
HTONSD	Home Team shots on targets Difference	HTSS (Shots On target taken) – HTSC (Shots On target conceded)
ATONSD	Away Team shots on targets Difference	ATSS (Shots On target taken) – ATSC (Shots On target conceded)
HTOFSD	Home team shots off targets Difference	HTSS (Shots Off target taken) – HTSC (Shots Off target conceded)
ATOFSD	Away Team shots off targets Difference	ATSS (Shots Off target taken) – ATSC (Shots Off target conceded)
HTYD	Home team yellow cards Difference	HTYC (Yellow cards received) – HTYG (Yellow cards opponent received)

ATYD	Away Team yellow cards Difference	ATYC (Yellow cards received) – ATYG (Yellow cards opponent received)
HTRD	Home team red cards Difference	HTYC (Red cards received) – HTYG (Red cards opponent received)
ATRD	Away Team red cards Difference	ATYC (Red cards received) – ATYG (Red cards opponent received)
HTCRD	Home team crosses made Difference	HCR1 (Crosses made) – HCR2 (Crosses the opponent made)
ATCRD	Away Team crosses made Difference	ACR1 (Crosses made) – ACR2 (Crosses the opponent made)
HTCOD	Home team corners Difference	HCO1 (Corners received) – HCO2 (Corners conceded)
ATCOD	Away Team corners Difference	ACO1 (Corners received) – ACO2 (Corners conceded)
HPO	Home team possession Difference	HPO1 (possession) – HPO2 (Opponent possession
APO	Away team possession Difference	APO1 (possession) – APO2 (Opponent possession

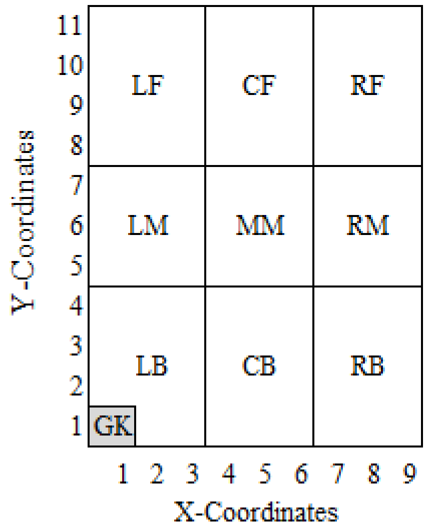
In order to answer the second sub-question, the team-statistics table was closely studied. As can be seen in Table 5, the team-statistics consisted of two columns for each variable, the Integer and the class. After viewing the columns we decided to use the integer values for each variable since they contain more detailed information compared to the class values. The data visualization showed that for the variable "Build up play Dribbling" more than half of the observations were missing so we decided to exclude this from the research. The remaining variables all have a highly significant correlation with the independent variable, however, the correlation strength is not as high as for the in-match statistics. All team-specific features are standardized before checking the correlations and controlling for multicollinearity.

Table 5: The description of the final team statistics in the dataset

Data Cleaning: Team Statistics (Top 5 Leagues)	
Variables	**Description**
build Up Play Speed	Home Team Speed of Build-up play
Build Up Play Passing	Home Team Passing quality of Build-up play
Chance Creation Passing	Home Team Chance creation passing quality
Chance Creation Crossing	Home Team Chance creation crossing quality
Chance Creation Shooting	Home Team Chance creation shooting quality
Defence Pressure	Home Team Defensive pressure high line
Defence Aggression	Home Team Aggression in defending
Defence Team Width	Home Team Width of the defensive shape
A_build Up Play Speed	Away team Speed of Build-up play
A_Build Up Play Passing	Away team Passing quality of Build-up play
A_Chance Creation Passing	Away team Chance creation passing quality
A_Chance Creation Crossing	Away team Chance creation crossing quality
A_Chance Creation Shooting	Away team Chance creation shooting quality
A_Defence Pressure	Away team Defensive pressure high line
A_Defence Aggression	Away team Aggression in defending
A_Defence Team Width	Away team Width of the defensive shape

In comparison to the team-statistics, the player-statistics studied in sub-question 3 required some more feature extraction before valuable insights could be drawn from them. The datasets “player attributes” was linked with the “match” dataset in order to retrieve all the necessary information from the dataset. There are several steps that need to be taken before the features in this dataset can be used for predicting the match result. Firstly, the positions of each player have to be generated from the X and Y coordinates available in the Match table. The X coordinate of each player denotes the place on the width of the field of a player ranging from 1-9, while the Y coordinate denotes the position of a player on the length of the field ranging from 1-11. Figure 3 displays the translation of the X and Y coordinates to real positions on the field. Each player number is not automatically linked to a position, so after creating dummies for all the player positions possible there are 61 positional dummies remaining.

Figure 3: Engineering new positional feature for the player-attributes.



Secondly, the ID from each player in a match has to be merged against the FIFA player dataset in order to retrieve the correct player statistics for each player in the match. This has to be done for each season as player statistics develop over the years, similar to the team statistics. At first, this will be done for the overall rating of a player, after which this will be done for all 35 player statistics we have available. The dataset resulting from this computation will have the player statistics for each player in a specific match. These statistics will be linked with the position of the player in order to get the value in one column, for example: “Home_Player_1_GK: 70 (rated)”. Whenever there are multiple players which play in the same position in one match, for example, two Central backs, these are averaged and one final value is taken for the “Home_Player_CB” in one match. It can also occur that there are no specific players playing in one of the 10 positions on the pitch. In order to avoid the loss of data, the mean average is taken to fill the Nan values in these rows for empty positions. The player statistics that eventually will be tested for correlation and multicollinearity will contain for each match the 35 player statistics (Table 3) of the Goalkeeper, Left-back, Central-back, Right-back, Left-midfielder, Central-midfielder, Right-midfielder, Left-forward, Central-forward and Right-forward. This will result in a total of 700 columns (10 home positions & 10 away positions for each of the 35 statistics). Lastly, Since a lot of the dependent variables have a correlation with each other, we decided to perform PCA on these variables. PCA is a good way to describe a great number of features with a smaller amount of principal components. At first, the goalkeeper statistics are removed from the ten outfield players and the outfield player statistics are excluded from the goalkeeper variable. Hereafter, the correlation is tested of all the variables remaining, the variables with a p-value greater than 0.05 are excluded from the research. Principal component analysis is performed on the remaining variables per position. The most important player statistic for match prediction is the overall rating since this is a summary of all the other statistics. A separate PCA analysis is done on the difference between the home team rating and away team rating per position. Principal components with a significance level below 95% are removed from the dataset. The final dependent variables that will answer the sub-question three can be seen in Table 6.

Table 6: The description of the final player statistics in the dataset

PCA	**Variables that are used in this PCA per position**
PC_RB1, PC_RB2, PC_RB3	RB Potential, RB crossing, RB finishing, RB heading accuracy, RB short passing, RB volleys, RB dribbling, RB curve, RB freekick accuracy, RB long passing, RB ball control, RB acceleration, RB sprint speed, RB agility, RB reactions, RB balance, RB shot power, RB jumping, RB stamina, RB strength, RB long shots, RB aggression, RB interceptions, RB positioning, RB vision, RB penalties, RB marking, RB standing tackle, RB sliding tackle
PC_LB1, PC_LB2, PC_LB3	LB Potential, LB Crossing, LB Finishing, LB Heading, LB Short passing, LB Volleys, LB Dribbling, LB Curve, LB Freekick, LB Long Passing, LB Ball Control, LB Acceleration, LB Sprint speed, LB Agility, LB Reactions, LB Balance, LB Shot power, LB Jumping, LB Stamina, LB Strength, LB Long shots, LB Aggression, LB Interceptions, LB Positioning, LB Vision, LB Penalties, LB Marking, LB Standing tackle, LB Sliding tackle
PC_CB1, PC_CB2, PC_CB3	CB Potential, CB Crossing, CB Finishing, CB Heading, CB Short passing, CB Volleys, CB Dribbling, CB Curve, CB Freekick, CB Long Passing, CB Ball Control, CB Acceleration, CB Sprint speed, CB Agility, CB Reactions, CB Balance, CB Shot power, CB Jumping, CB Stamina, CB Strength, CB Long shots, CB Aggression, CB Interceptions, CB Positioning, CB Vision, CB Penalties, CB Marking, CB Standing tackle, CB Sliding tackle
PC_RM1, PC_RM2, PC_RM3	RM Potential, RM Crossing, RM Finishing, RM Heading, RM Short passing, RM Volleys, RM Dribbling, RM Curve, RM Freekick, RM Long Passing, RM Ball Control, RM Acceleration, RM Sprint speed, RM Agility, RM Reactions, RM Balance, RM Shot power, RM Jumping, RM Stamina, RM Strength, RM Long shots, RM Interceptions, RM Positioning, RM Vision, RM Penalties
PC_LM1,	LM Potential, LM Crossing, LM Finishing, LM Heading, LM Short passing, LM Volleys, LM Dribbling, LM Curve, LM Freekick, LM Long Passing, LM Ball Control, LM Acceleration, LM Sprint speed, LM Agility, LM Reactions, LM Balance, LM Shot power,

PC_LM2, PC_LM3	LM Jumping, LM Stamina, LM Strength, LM Long shots, LM Aggression, LM Interceptions, LM Positioning, LM Vision, LM Penalties, LM Marking, LM Standing tackle, LM Sliding tackle
PC_MM1, PC_MM2, PC_MM3	MM Potential, MM Crossing, MM Finishing, MM Heading, MM Short passing, MM Volleys, MM Dribbling, MM Curve, MM Freekick, MM Long Passing, MM Ball Control, MM Acceleration, MM Sprint speed, MM Agility, MM Reactions, MM Balance, MM Shot power, MM Jumping, MM Stamina, MM Strength, MM Long shots, MM Aggression, MM Interceptions, MM Positioning, MM Vision, MM Penalties, MM Marking, MM Standing tackle, MM Sliding tackle
PC_RF1, PC_RF2, PC_RF3	RF Potential, RF Crossing, RF Finishing, RF Heading, RF Short passing, RF Volleys, RF Dribbling, RF Curve, RF Freekick, RF Long Passing, RF Ball Control, RF Acceleration, RF Sprint speed, RF Agility, RF Reactions, RF Balance, RF Shot power, RF Jumping, RF Stamina, RF Strength, RF Long shots, RF Aggression, RF Interceptions, RF Positioning, RF Vision, RF Penalties, RF Marking
PC_LF1, PC_LF2, PC_LF3	LF Potential, LF Crossing, LF Finishing, LF Heading, LF Short passing, LF Volleys, LF Dribbling, LF Curve, LF Freekick, LF Long Passing, LF Ball Control, LF Acceleration, LF Sprint speed, LF Agility, LF Reactions, LF Balance, LF Shot power, LF Jumping, LF Stamina, LF Long shots, LF Interceptions, LF Positioning, LF Vision, LF Penalties
PC_CF1, PC_CF2, PC_CF3	CF Potential, CF Crossing, CF Finishing, CF Heading, CF Short passing, CF Volleys, CF Dribbling, CF Curve, CF Freekick, CF Long Passing, CF Ball Control, CF Acceleration, CF Sprint speed, CF Agility, CF Reactions, CF Balance, CF Shot power, CF Jumping, CF Stamina, CF Strength, CF Long shots, CF Aggression, CF Interceptions, CF Positioning, CF Vision, CF Penalties, CF Marking, CF Standing tackle
PC_Diff1, PC_Diff2, PC_Diff3, PC_Diff4	GK Diff, RB Diff, LB Diff, CB Diff, LM Diff, MM Diff, RM Diff, LF Diff, CF Diff, RF Diff

Evaluation method and final models

Three different classifiers will be tested to see which performs the best using our set of features. The Sci-kit learn package in python includes the first two classifiers, Random forest and Logistic regression. The last classifier that will be used is the XgBoost classifier. The open-source framework of (Chen & Guestrin, 2016) has been used for implementing gradient boost on our feature set. The final evaluation will be based on the performance of the final prediction model. The final model of this research will be evaluated against benchmark predictive models in the field of football match prediction. If the final model obtains accuracy and metrics that are superior to previous research their metrics then the final model has highlighted some essential attributes which influence the match outcome. The final model will work like the Formula 1,2 and 3 described below. The formulas below include the sole effect of the match statistics, team statistics and player statistics on the match outcome. The betas represent the correlation they have with the independent variable, however, the correlation between the dependent variables will reduce the prediction accuracy. All feature sets used in the three sub-questions are displayed in the formulas below. The creation of these features will be explained in the next section.

Formula 1: Presentation of match-statistics features in prediction model

$$
\begin{aligned}
&Match\ Outcome_{Match-statistics} \\
&\quad = \beta1 * DiffPts + \beta2 * HTGD + \beta3 * DiffFormPts + \beta4 * HTONSD + \beta5 \\
&\quad * HTOFSD + \beta6 * HTCOD + \beta7 * HTCRD + \beta8 * HTP + \beta9 * ATYD + \beta10 \\
&\quad * ATRD + \beta11 * HTRD + \beta12 * HTYD + \beta13 * ATP + \beta14 * ATCRD + \beta15 \\
&\quad * ATCOD + \beta16 * ATONSD + \beta17 * ATCOD + \beta18 * ATONSD + \beta19 * ATOFSD \\
&\quad + \beta20 * ATGD + \beta21 * DiffLP
\end{aligned}
$$

Formula 2: Presentation of team-statistics features in prediction model

$$
\begin{aligned}
&Match\ Outcome_{Team-statistics} \\
&\quad = \beta1 * Home\ Build\ up\ play\ speed + \beta2 * Home\ Build\ up\ play\ passing + \beta3 \\
&\quad * Home\ Chance\ creation\ passing + \beta4 * Home\ Chance\ creation\ crossing + \beta5 \\
&\quad * Home\ Chance\ creation\ shooting + \beta6 * Home\ Defence\ pressure + \beta7 \\
&\quad * Home\ Defence\ Aggression + \beta8 * Home\ Defence\ Team\ width + \beta9 \\
&\quad * Away\ Build\ up\ play\ speed + \beta10 * Away\ Build\ up\ play\ passing + \beta11 \\
&\quad * Away\ Chance\ creation\ passing + \beta12 * Away\ Chance\ creation\ crossing \\
&\quad + \beta13 * Away\ Chance\ creation\ shooting + \beta14 * Away\ Defence\ pressure \\
&\quad + \beta15 * Away\ Defence\ Aggression + \beta16 * Away\ Defence\ Team\ width
\end{aligned}
$$

Formula 3: Presentation of player-statistics features in prediction model

$$
\begin{aligned}
&Match\ Outcome_{Player-statistics} \\
&\quad = \beta1 * PC_{RB1} + \beta2 * PC_{RB2} + \beta3 * PC_{RB3} + \beta4 * PC_{LB1} + \beta5 * PC_{LB2} + \beta6 \\
&\quad * PC_{CB1} + \beta7 * PC_{CB2} + \beta8 * PC_{LM1} + \beta9 * PC_{LM2} + \beta10 * PC_{MM1} + \beta11 * PC_{MM2} \\
&\quad + \beta12 * PC_{MM3} + \beta13 * PC_{RM1} + \beta14 * PC_{RM2} + \beta15 * PC_{RM3} + \beta16 * PC_{LF1} \\
&\quad + \beta17 * PC_{LF2} + \beta18 * PC_{LF3} + \beta19 * PC_{A_{RB_1}} + \beta20 * PC_{A_{RB_2}} + \beta21 * PC_{A_{RB3}} \\
&\quad + \beta22 * PC_{A_{LB_1}} + \beta23 * PC_{A_{LB_2}} + \beta24 * PC_{A_{CB_1}} + \beta5 * PC_{A_{CB_2}} + \beta26 * PC_{A_{LM1}} \\
&\quad + \beta27 * PC_{A_{LM_2}} + \beta28 * PC_{A_{MM_1}} + \beta29 * PC_{A_{MM_2}} + \beta30 * PC_{A_{RM_1}} + \beta31 * PC_{A_{LF1}} \\
&\quad + \beta32 * PC_{A_{LF_2}} + \beta33 * PC_{A_{CF_1}} + \beta34 * PC_{A_{RF_1}} + \beta35 * PC_{A_{RF_2}} + \beta36 * PC_{A_{RF3}} \\
&\quad + \beta37 * PC_{A_{GK_1}} + \beta38 * PC_{A_{GK_3}} + \beta39 * PC_{GK1} + \beta40 * PC_{GK2} + \beta41 * PC_{GK3} \\
&\quad + \beta44 * PC_{diff1} + \beta44 * PC_{diff2} + \beta44 * PC_{diff3}
\end{aligned}
$$

5. Results

This chapter describes the results of the three different analysis that are conducted in order to be able to answer the research questions. The first section will discuss the effect of match statistics on the outcome of football matches. The second section will discuss this effect for team-specific attributes. Whereafter, the effect player-specific features on the match outcome will be displayed in section three. The prediction accuracy in each sub-question will be displayed for the 5 top leagues studied separately and combined. Lastly, a final model is displayed where the player statistics and match statistics dataset is merged

Match-statistics

This analysis is conducted to be able to answer the first sub-question:

SQ 1: Which historical match features correlate with the outcome of a football match ?

In order to view the most relevant features for the match predictor, the Pearson correlation coefficient was used to view the correlation between the dependent and the independent variable. This correlation coefficient was also used in the research of Kampakis et al. (2015), they concluded that this was the best feature selection measurement to use. The most important features can be seen in the correlation table for match statistics (Table 7).

Table 7: Correlation matrix of the set of Match-statistics features

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on in-match statistics and team rankings. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

Data Cleaning: Match Statistics (Top 5 Leagues)				
Variables	**Description**	**Correlation**	**P-Value**	**VIF**
DiffPts	Difference in League points this season	0.28	0.00	Inf
HTGD	Home Team Goal Difference	0.21	0.00	4.63
DIffFormPts	Difference in points in the last matches	0.18	0.00	2.17
HTONSD	Home Team shots on targets Difference	0.16	0.00	3.49
HTOFSD	Home team shots off targets Difference	0.15	0.00	3.05
HTCOD	Home team corners Difference	0.15	0.00	3.83
HTCRD	Home team crosses made Difference	0.14	0.00	2.75
HTP	Home team points in season	0.13	0.00	Inf
ATYD	Away Team yellow cards Difference	0.09	0.00	1.21
ATRD	Away team red cards Difference	0.03	0.00	1.03
HTRD	Home team red cards Difference	-0.02	0.00	1.04

HTYD	Home team yellow cards Difference	-0.11	0.00	1.21
ATP	Away team points in season	-0.13	0.00	Inf
ATCRD	Away Team crosses made Difference	-0.14	0.00	2.74
ATCOD	Away Team corners Difference	-0.15	0.00	3.84
ATONSD	Away Team shots on targets Difference	-0.16	0.00	3.52
ATOFSD	Away Team shots off targets Difference	-0.16	0.00	3.08
ATGD	Away Team Goal Difference	-0.21	0.00	4.56
DiffLP	Difference in League position last season	-0.29	0.00	1.49

This table displays the correlation between the final features extracted from the dataset and the match outcome. In order to use the Pearson correlation test, the FTR variable was briefly changed from (Home, Draw, Away) to (3 Points, 1 Point, 0 Points). As can be seen in the table, all features have a p-value of 0.00 which means that we can state with the confidence of over 99% that these features correlate with the match result. The most important variables for our first predictor seem to be the Difference in league points (DiffPts) this season and Difference in League position (DiffLP) last season. Since all these features are significantly correlated with the independent variable and multicollinearity is not present in this dataset, all features are used in the prediction model.

The classifiers that are used in this research to predict the match outcome are Random forest, Gradient Boost and Logistic regression. In previous research, the random forest classifier often outperformed other classifiers. This is not in line with this research, where Table 8 displays that the best predicting classifier in four out of 6 datasets was the Gradient boost classifier. As can be seen in the confusion matrix in Table 8, the model obtains the highest accuracy when predicting home victories.

Table 8: Confusion matrix of the predictor based on match statistics

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on in match statistics and league rankings. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

	England (Logistic Regression)				Germany (Random Forest)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	102	5	51	0.65	52	10	73	0.39
D Actual	61	6	77	0.04	30	13	65	0.12
H Actual	64	17	177	0.69	51	4	138	0.68
	Spain (XGBoost)				Italy (XGBoost)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	66	8	82	0.42	76	8	63	0.52
D Actual	22	1	105	0.01	40	9	96	0.06
H Actual	29	2	243	0.88	30	12	212	0.83
	France (XGBoost)				Top Five Leagues grouped (XGBoost)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	42	14	94	0.28	302	2	443	0.40
D Actual	33	18	107	0.11	173	4	505	0.01
H Actual	42	15	195	0.77	174	5	1064	0.86

These accuracies in the six datasets vary from 0.68 in Germany to 0.88 in Spain. The model produces the worst results when predicting draws. Similar to previous research, it remains difficult to predict a draw because the model will often lean towards a win or a loss. In the combined dataset of the top five leagues, only 1% of the actuals draws are predicted correctly. As said before, this is mainly due to the overpredicting of losses and wins since only 11 out of the total 2672 matches in the test set were predicted to end in a draw.

The total accuracy of the final model on the six datasets is displayed above in Table 9. The highest accuracy is obtained when predicting the Spanish league using an XGBoost classifier. This accuracy of 0.56 is comparable to the best-obtained result from previous researches studied. The research of Baboota & Kaur (2018) predicted the English premier league using a random forest classifier with an accuracy of 0.57. The worst accuracy is obtained for the German league with an average accuracy of 0.45. The combined dataset performed similarly to the country-specific datasets.

Table 9: Accuracy scores for the predictor based on match statistics

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on in-match statistics and team rankings. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes for each of the classifiers. All the features that were used in these predictors were all significant at 1% level.

	England	Germany	Spain	Italy	France	Top 5
Random Forest Classifier	0.50	0.46	0.53	0.53	0.45	0.49
XGBoost Classifier	0.49	0.45	0.56	0.54	0.46	0.51
Logistic Regression Classifier	0.51	0.44	0.55	0.51	0.45	0.50
N Test Size	560	446	560	546	560	2672

Team-attributes

This analysis is conducted to be able to answer the second sub-question:

SQ 2: Which team attributes correlate with the outcome of a football match ?

The analysis of the correlations displayed that all features that will be used are significant and that multicollinearity is not an issue in this dataset. The features which have the most positive effect on the match outcome is the pressure of the defence. While the passing quality of build-up play has a surprisingly negative effect on the match outcome. A reason for this could be that build up play passing doesn’t directly lead to chances so this does not positively influence a match outcome. All correlation coefficients for the team attributes are displayed in Table 10 below.

Table 10: Correlation matrix of the set of Team-statistics features

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on in-match statistics and team rankings. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

Data Cleaning: Team Statistics (Top 5 Leagues)				
Variables	**Description**	**Correlation**	**P-Value**	**VIF**
Defence Pressure	Home Team Defensive pressure high line	0.07	0.00	1.5
Chance Creation Shooting	Home Team Chance creation shooting quality	0.04	0.00	1.1
Chance Creation Passing	Home Team Chance creation passing quality	0.02	0.00	1.3
Defence Aggression	Home Team Aggression in defending	0.02	0.00	1.2
Defence Team Width	Home Team Width of the defensive shape	0.02	0.00	1.3
A_Chance Creation Crossing	Away team Chance creation crossing quality	0.01	0.00	1.1
A_Defence Pressure	Away team Defensive pressure high line	0.01	0.00	1.5
A_Chance Creation Shooting	Away team Chance creation shooting quality	-	0.00	1.1
A_Defence Aggression	Away team Aggression in defending	-	0.00	1.2
A_Build Up Play Passing	Away team Passing quality of Build-up play	-0.01	0.00	1.3
Chance Creation Crossing	Home Team Chance creation crossing quality	-0.02	0.00	1.2
build Up Play Speed	Home Team Speed of Build-up play	-0.03	0.00	1.3
A_build Up Play Speed	Away team Speed of Build-up play	-0.03	0.00	1.3
A_Chance Creation Passing	Away team Chance creation passing quality	-0.03	0.00	1.2
Build Up Play Passing	Home Team Passing quality of Build-up play	-0.09	0.00	1.4

The predicting model based solely on team attributes does not obtain the best results for our dataset. As can be seen in Table 11, the model overpredicts a match win tremendously.

Table 11: Confusion matrix of the predictor based on team statistics

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on team attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

	England (XGBoost)				Germany (XGBoost)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	50	15	65	0.38	51	7	54	0.46
D Actual	31	13	72	0.11	36	2	51	0.02
H Actual	51	26	133	0.63	35	13	117	0.71
	Spain (XGBoost)				Italy (XGBoost)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	21	4	101	0.17	28	10	84	0.23
D Actual	13	8	87	0.07	27	8	86	0.07
H Actual	27	10	185	0.83	30	19	158	0.76
	France (XGBoost)				Top Five Leagues grouped (XGBoost)			
	A Predict	D Predict	H Predict	Accuracy	A Predict	D Predict	H Predict	Accuracy
A Actual	16	17	90	0.13	44	7	563	0.07
D Actual	19	17	92	0.13	35	5	521	0.01
H Actual	17	22	166	0.81	40	10	961	0.95

This results in high accuracy for predicting a match win, however, the prediction of a loss or a draw obtains a very low accuracy. In the combined dataset of the top five leagues, 2045 out of the total 2186 matches in the test set were predicted to end in a home win. This is roughly 94%, where on average the home team wins approximately 45% of all matches.

Table 12 below displays the total accuracy for the prediction model based on team attributes. Similar to the match-statistics, the best accuracy is obtained when using an XGBoost classifier on the Spanish football competition. An accuracy of 47% is obtained which is still better than predicting the match outcome based on chance, which results in a 33% accuracy. However, this 0.47 is 0.09 lower than the accuracy obtained in section one and 0.10 lower than the research of Baboota & Kaur. (2018). The predicted accuracy on the combined dataset is similar to the average home wins percentage of this dataset (0.46).

Table 12: Team Statistics as a predictor for Match results

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on Team-specific attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes for each of the classifiers. All the features that were used in these predictors were all significant at 1% level.

	England	Germany	Spain	Italy	France	Top 5
Random Forest Classifier	0.40	0.39	0.40	0.38	0.39	0.40
XGBoost Classifier	0.43	0.46	0.47	0.41	0.44	0.46
Logistic Regression Classifier	0.42	0.43	0.46	0.42	0.41	0.44
N Test Size	456	366	456	450	456	2186

Player-attributes

This analysis is conducted to be able to answer the third sub-question:

SQ 3: Which Player attributes correlate with the outcome of a football match ?

In order to study this dataset, a large feature extraction and selection process has been done. After generating all the features and performing PCA on these features to reduce the dimensionality of the features, three principal components were made for each position. Three components were chosen because for each position there will be defensive, overall and offensive statistics. The components were checked for significance by using the Pearson correlation coefficient. The significant features in this dataset are listed below in Table 13. All features with a Variance-Inflation factor above 10 are normally removed from the dataset, however, due to the high correlation of the PC_diff1 feature with the independent variable, this feature was not excluded.

Table 13: Correlation matrix of the set of player-statistics features

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on player-attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 5% level.

Data Cleaning: Player Statistics (Top 5 Leagues)				
Variables	**Description**	**Correlation**	**P-Value**	**VIF**
PC_diff1	Principal Component 1 Difference Overall Rating	0.35	0	30.3
PC_CB1	Principal Component 1 Central-back (Home Team)	0.21	0	1.89
PC_ A_CF1	Principal Component 1 Central-forward (Away Team)	0.18	0	2.97
PC_ A_MM1	Principal Component 1 Mid-mid (Away Team)	0.15	0	2.16
PC_A_RB1	Principal Component 1 Right-back (Away Team)	0.14	0	2.27
PC_ A_LB1	Principal Component 1 Left-back (Away Team)	0.14	0	2.84
PC_ A_RF1	Principal Component 1 Right-forward (Away Team)	0.14	0	2.88
PC_A_GK1	Principal Component 1 Goalkeeper (Away Team)	0.14	0	7.12
PC_ A_LF1	Principal Component 1 Left-forward (Away Team)	0.13	0	1.71
PC_ A_RM1	Principal Component 1 Right-mid (Away Team)	0.11	0	3.06
PC_ A_LM1	Principal Component 1 Left-mid (Away Team)	0.1	0	2.67
PC_ A_CB2	Principal Component 2 Central -back (Away Team)	0.09	0	1.42
PC_ A_RB2	Principal Component 2 Right-back (Away Team)	0.08	0	1.48
PC_ A_LB2	Principal Component 2 Left-back (Away Team)	0.06	0	2.01
PC_ A_MM2	Principal Component 2 Mid-mid (Away Team)	0.05	0	1.22
PC_ A_RB3	Principal Component 3 Right-back (Away Team)	0.04	0	1.08
PC_ A_LF2	Principal Component 2 Left-forward (Away Team)	0.04	0	1.04
PC_diff4	Principal Component 4 Difference Overall Rating	0.04	0	8.5
PC_GK2	Principal Component 2 Goalkeeper (Home Team)	0.03	0	1.11
PC_GK3	Principal Component 3 Goalkeeper (Home Team)	0.03	0	1.04
PC_ A_RF3	Principal Component 3 Right -forward (Away Team)	0.03	0	1.03
PC_RM2	Principal Component 2 Right -mid (Home Team)	0.02	0.03	1.19
PC_ A_LM2	Principal Component 2 Left-mid (Away Team)	0.02	0.02	1.25
PC_ A_RF2	Principal Component 2 Right -forward (Away Team)	0.02	0.01	1.04
PC_RB3	Principal Component 3 Right-back (Home Team)	-0.02	0.01	1.09
PC_MM3	Principal Component 3 Mid-mid (Home Team)	-0.02	0.02	1.13
PC_RM3	Principal Component 3 Right -mid (Home Team)	-0.02	0.02	1.11
PC_LF1	Principal Component 2 Left-forward (Home Team)	-0.02	0.04	1.05
PC_RF3	Principal Component 3 Right -forward (Home Team)	-0.02	0.04	1.04
PC_A_GK3	Principal Component 3 Goalkeeper (Away Team)	-0.02	0.01	1.04
PC_diff2	Principal Component 2 Difference Overall Rating	-0.02	0.01	5.94
PC_LM2	Principal Component 2 Left-mid (Home Team)	-0.03	0	1.29
PC_MM2	Principal Component 2 Mid-mid (Home Team)	-0.04	0	1.25

PC_RB2	Principal Component 2 Right-back (Home Team)	-0.07	0	1.51
PC_LB2	Principal Component 2 Left-back (Home Team)	-0.07	0	2.02
PC_CB2	Principal Component 2 Central -back (Home Team)	-0.07	0	1.47
PC_RM1	Principal Component 1 Right-mid (Home Team)	-0.12	0	3.11
PC_LM1	Principal Component 1 Left-mid (Home Team)	-0.13	0	2.64
PC_LF1	Principal Component 1 Left-forward (Home Team)	-0.13	0	1.76
PC_GK1	Principal Component 1 Goalkeeper (Home Team)	-0.14	0	7.22
PC_LB1	Principal Component 1 Left-back (Home Team)	-0.15	0	2.78
PC_RB1	Principal Component 1 Right-back (Home Team)	-0.16	0	2.31
PC_RF1	Principal Component 1 Right-forward (Home Team)	-0.16	0	3.11
PC_MM1	Principal Component 1 Mid-mid (Home Team)	-0.17	0	2.15
PC_ A_CB1	Principal Component 1 Central-back (Away Team)	-0.19	0	1.93
PC_CF1	Principal Component 1 Central-forward (Home Team)	-0.21	0	3.00

The final model based on player statistics outperformed the model based on team attributes. Especially when looking at the prediction of an Away win (Table 14). This accuracy has significantly increased in comparison to the previous model. The accuracy of predicting a home win is high in all country-specific datasets, varying from an accuracy of 0.71 in England to 0.85 in Spain. The interesting result is that whenever the model is trained on more data, as is the case for the combined dataset, the predictor more often predicts home wins with an accuracy of 0.87. This means that whenever the data availability increases in this dataset, the prediction of a home win grows. Similar to sub-question one, predicting a draw is still difficult for the model, with accuracies ranging from 0.00 in Spain to 0.12 in England and France.

Table 14: Confusion matrix of the predictor based on Player statistics

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on player-specific attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

	England (Logistic Regression)				Germany (XGBoost)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	105	11	105	11	105	11	105	11
D Actual	64	18	64	18	64	18	64	18
H Actual	56	21	56	21	56	21	56	21
	Spain (XGBoost)				Italy (XGBoost)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	62	6	62	6	62	6	62	6
D Actual	27	0	27	0	27	0	27	0
H Actual	37	2	37	2	37	2	37	2
	France (Logistic Regression)				Top Five Leagues grouped (XGBoost)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	62	12	62	12	62	12	62	12
D Actual	40	19	40	19	40	19	40	19
H Actual	30	24	30	24	30	24	30	24

The accuracies obtained for the player statistics dataset are comparable to previous researches. Especially, the results of the combined dataset. The accuracy of 0.50-0.52 is in line with the results from Ulmer et al. (2013), Constatinou et al. (2012) and Gomes et al. (2015) (0.50-0.51).

Table 15: Player Statistics as a predictor for Match results

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on Player-specific attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes for each of the classifiers. All the features that were used in these predictors were all significant at 5% level.

	England	Germany	Spain	Italy	France	Top 5
Random Forest Classifier	0.52	0.47	0.52	0.50	0.47	0.50
XGBoost Classifier	0.51	0.49	0.53	0.50	0.50	0.52
Logistic Regression Classifier	0.54	0.45	0.51	0.51	0.49	0.50
N Test Size	592	474	540	548	572	2730

The prediction model for the English premier league using a Logistic regression classifier obtains the highest accuracy (Table 15). The XGBoost classifier obtained the highest accuracy in three out of the other five datasets, making it the classifier that has performed the best in this research.

Final model

This analysis is conducted to be able to answer first research question:

RQ: "To what extent can investing in high-depth data analysis be beneficial for the prediction of match results".?

Table 16: Final Statistics as a predictor for Match results

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on Player-specific and match-specific attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes, varying from Home win, Away win and Draw. The outcomes of the classifier that obtained the best accuracy are presented. All the features that were used in these predictors were all significant at 1% level.

	England (Random Forest)				Germany (XGBoost)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	72	15	72	15	72	15	72	15
D Actual	45	17	45	17	45	17	45	17
H Actual	41	13	41	13	41	13	41	13
	Spain (Random Forest)				Italy (Logistic Regression)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	49	13	49	13	49	13	49	13
D Actual	24	11	24	11	24	11	24	11
H Actual	26	19	26	19	26	19	26	19
	France (XGBoost)				Top Five Leagues grouped (XGBoost)			
	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict	A Predict	D Predict
A Actual	65	11	65	11	65	11	65	11
D Actual	33	17	33	17	33	17	33	17
H Actual	34	20	34	20	34	20	34	20

After reviewing the results of sub-questions 1, 2 and 3. The final model was built by joining the models which provided good accuracies (Table 16). Therefore, the model based on team-attributes was excluded from this final model. This final model still accurately predicts if a home team wins the match. The combined dataset obtains an accuracy of 0.87 on the home win prediction. However, there has been little improvement to the prediction of the draw and away win.

Table 17: Final Statistics as a predictor for Match results

This table presents an overview of the effect of the predictor on football matches in various European leagues . This predictor is fully based on Player-specific and match-specific attributes. The leagues that were studied are the top 5 leagues in Europe: England, Germany, Spain, Italy and France. Below figures present the correctly predicted outcomes for each of the classifiers. All the features that were used in these predictors were all significant at 5% level.

	England	Germany	Spain	Italy	France	Top 5
Random Forest Classifier	0.52	0.49	0.52	0.50	0.49	0.51
XGBoost Classifier	0.49	0.53	0.51	0.48	0.50	0.52
Logistic Regression Classifier	0.50	0.50	0.52	0.51	0.47	0.51
N Test Size	540	434	498	548	572	2730

Combining the match-based and player-based models do not improve most of the accuracies in this study, as can be seen in Table 17 below. The highest accuracy is obtained by predicting the German League with an XGBoost classifier. This is surprising compared to the results from the sub-questions where the German league predictor performed as one of the worst. This predictor of the German league obtained an accuracy of 0.53, which is the highest accuracy obtained comparing all the predictors. However, the predictor of most other leagues obtained similar or better results when only using the match-statistics or player-statistics.

6. Discussion

The main goal of this research was to find and create features which have a positive influence on predicting the match outcome of a football match. After these features were created, a model was made to test if the prediction model could outperform the benchmark and previous researchers who had the same objective. Three main pieces of researches described in the literature review are used for comparison with our model in this discussion. Firstly, Baboota & Kaur (2018) since this is the most recent research on this topic. The research of Tax et al. (2015) is the second research used for comparison, due to the fact that they study a different league compared to the other researchers. Lastly, Gomes et al. (2015) was selected for this discussion, since they used a similar set of features in their analysis compared with our match-statistics analysis.

In order to be able to generate a good predictor, new features had to be created from the existing data. As there was a huge amount of features available in the dataset, the decision was made to split the research question into three sub-questions. This decision was primarily made because a lot of the features have a high correlation with each other. Having all these features in different dataset solves the problem of multicollinearity and opens up the possibility to view what subset of the data correlates the most with a match outcome.

The first model based on match statistics proved to be the best predictor of the match outcome in our research. The main features that contributed to this positive outcome were the difference in League points in the corresponding season and the league position in last season. Furthermore, variables like shots, corners, crosses, and cards also improved the accuracy of the model. These results were expected when studying the previous research of Gomes et al. (2015). They obtained a similar accuracy when predicting the match outcome of the English Premier league using a Support vector machine compared to our Logistic regression classifier (0.51). However, this was not our best result for the predictor based on match statistics. The best-obtained accuracy was 0.56 for the Spanish league using gradient boost. These results are similar to the results Baboota & Kaur (2018) obtained for their predictor using gradient boost. Baboota & Kaur (2018) only studied the English premier league, so it can be argued that when they would test their model on the Spanish league, their accuracy would be even higher than the 0.57 obtained in their research. The difference between the research of Baboota & Kaur (2018) and this research is the fact that they used 11 seasons of Premier League football compared to 8 seasons in our dataset. By gathering more data they could have trained their model better which lead to higher accuracy in the English premier league. The research of Tax et al. (2015) predicted match outcomes for the Dutch league. They obtained an accuracy of 54% by using random forest on a publicly available dataset. The use of Random Forest as a classifier for match statistics seemed to predict the match outcome above average for Spain and Italy. Predicting the match outcome with the random forest classifier in these countries led to an accuracy of 53% in both countries. These results are comparable with the results of Tax et al. (2015). However, for predicting the German and French leagues, the random forest classifier could not obtain those high accuracies. As can be seen in the result section, the German and French league predictors underperform in most models. An explanation for this unpredictability in these competitions could be that the clubs which play in these leagues are more evenly matched which makes it more difficult to predict match outcomes.

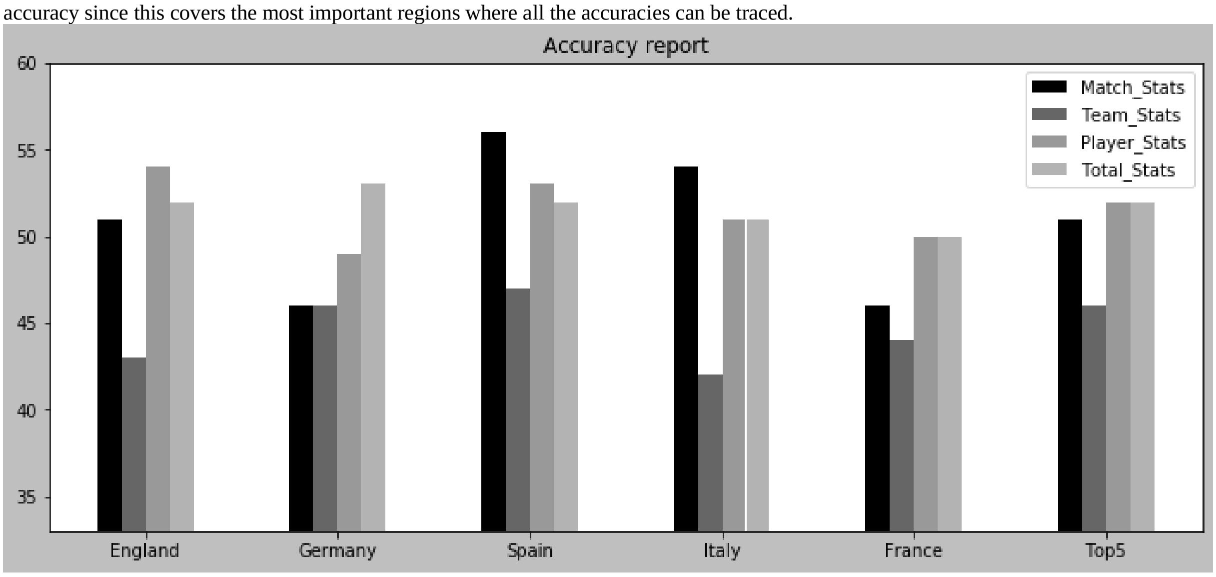
The second model built in this research was based on team attributes, these team attributes that were used are not similar to the features used in previous research. Therefore it was difficult to estimate how well the predictor based on these team attributes would perform. After seeing the results it can be said that the team attributes in this research are not a good predictor for the outcome of a match as they tend to overpredict a home win.

The third sub-question of this research could be answered using various models based on player attributes. In this research, we decided to use features that could show the strength of a player on a certain position on the pitch. PCA was performed per player to make sure we limit the data loss and all significant variables are taken into account. Next to this, the model also included one differential feature of the overall player rating per position. The thought behind these position-based ratings is that the important attributes per position could be studied and that these would all be included in the research. This approach is partly similar to the approach of Baboota & Kaur (2018). Their research included average ratings for the attack, midfield, and defense. So, in comparison with Baboota & Kaur (2018), our research has a more detailed view because we also consider the position on the width of the pitch. Baboota & Kaur (2018) did not perform a separate analysis on these player statistics which makes it difficult to compare these results. However, their final model including all features outperforms our model which could argue that a less detailed view could improve the accuracy.

Lastly, the final model is made by joining the dependent variables from the player-attributes and the match-statistics. Joining these features only improves the accuracy of the match outcome prediction in the German League. An explanation for this lack of added accuracy is due to the fact that most information from the player statistics is already captured in the match statistics indirectly. Figure 4 below displays a summary of the best accuracies obtained for each model in each region. As can be seen, all models outperform the 33% that accuracy is achieved without using a model. However, most models do not reach the 58% accuracy that was obtained by Baboota & Kaur (2018).

Figure 4: Accuracy report per country and model

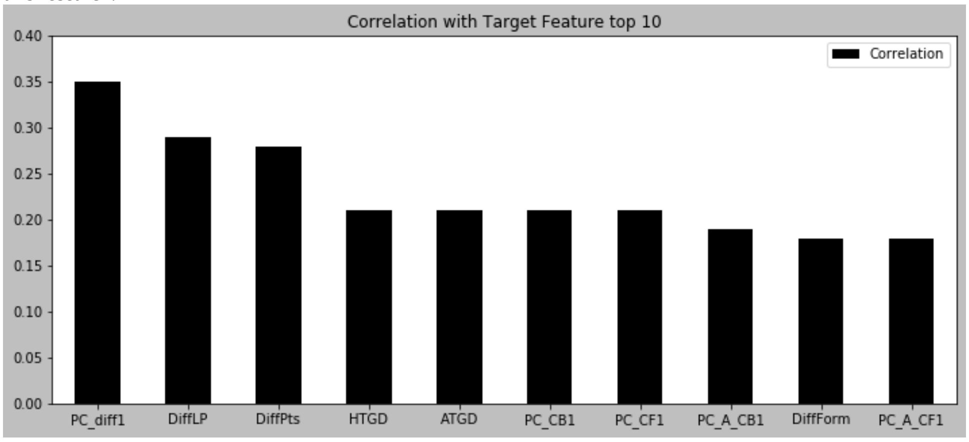
This figure presents an overview of the best accuracy obtained for each model in each country. Y-axis varies from 33% accuracy till 60% accuracy since this covers the most important regions where all the accuracies can be traced.



Next to outperforming previous researches and the benchmark, another important part of this research was focussed on discovering features that have a high correlation with the match outcome. Figure 5 below displays the top 10 features originating from the three models that were built. Six out of these ten features were used in the prediction model based on match-statistics, while the other four originate from the model based on player-statistics.

Figure 5: Top 10 features that correlate with the independent variable.

This figure presents an overview of the best features in this research. These top 10 features are drawn from the three models that were built in this research.



The correlations in this figure vary from 0.18 to 0.35, which are really high and significant correlations with the independent variable. The three most important features are the first principal component of the Difference in rating between the home and away players, the difference in league points and lastly the difference in league rankings.

7. Conclusion

This research addressed the following main question: "*To what extent can investing in high-depth data analysis be beneficial for the prediction of match results".* The upcoming sections describe the answers to the sub-questions first, after which the final research question will be answered.

***SQ1**: Do historical match related events help to predict the match outcome of a football match ?*

Previous research has been dominantly focussed on predicting match outcomes using these historical match related events. Similar to previous research on match statistics, the predictor based on match statistics built in this research also produces positive results. The predictor outperforms the benchmark of chance (33%) with approximately 20%. Next to this, the predictor also outperforms most previous researches done on this topic. Therefore, it can be concluded that the match related events help to predict the outcome of a football match.

***SQ2**: Does Team-specific data help to predict the match outcome of a football match ?*

Limited research has been done on team-specific attributes predicting the outcome of a football match. After viewing accuracies varying from 0.38 to 0.47, it can be concluded that the team attributes are not the best predictor of match results. The dependent variables used in this model overvalued the home win and therefore the results did improve the benchmark of chance, however, they did not outperform other studies based on match statistics.

***SQ3**: Does Player-specific data help to predict the match outcome of a football match ?*

The player-statistics contain a lot of valuable information on the strength of the total team. The results obtained using these statistics were close to previous researches. An interesting conclusion is that for the English premier league, the player statistics model outperformed the match statistics model with 3%. However, since the raw player specific attributes can have a high correlation with each other, an extensive feature selection and feature extraction process is necessary when using these attributes in the model.

The final research question "*To what extent can investing in high-depth data analysis be beneficial for the prediction of match results"* is a brief summary of the answer to the sub-questions. Investing in high-depth data analysis can be beneficial for a club in order to see the features that correlate with a positive match outcome. These features could be explored in more detail by clubs or researchers to gain additional insights into these features. Next to this, the match statistics model could help a team with their strategy, while the player-statistics model is useful for the acquisition of new players. This concludes that investing in high-depth data analysis can be beneficial for the prediction of match results.

8. Future Work and Limitations

As said before in the introduction, football is a sport that gets more popular every year. With the rise of big data, more and more data will be available. This new data could mean a breakthrough in the research field of sports prediction. So this type of research will have to be repeated on a regular basis with the newly available data.

The main difficulty in accurately predicting the outcome of a football match is the prediction of the game ending in a draw. Currently, in most models, the draw variable is underperforming massively. This is due to the fact that if the model leans a bit towards one team being better, they will assign a win to this team. In the future, it could be incorporated in the model that if the prediction for a win or loss is not strong enough, the model will choose a draw. Next to this draw correction in the model, an extensive analysis of what drives the match in ending in a draw would be valuable.

Additionally, it would be interesting to see these predictors tested on less popular leagues. This data is also available on Kaggle. However, due to time constraints, this research decided to focus solely on the top five leagues. Lastly, Various researchers have tried to predict the outcome in order to beat the odds provided by bookmakers so their model will be profitable. This has been a really difficult task, even with the high accuracies obtained by Baboota & Kaur (2015) a profit was not generated. It could be interesting to look at the prediction of other match statistics. There is the possibility of placing bets on the number of corners that will take place or the number of cards that will be given. These statistics could be more predictable than the match outcome and this opens up possibilities for a profitable model.

The limitations of this research are quite in line with the recommendations for future research. One limitation of this research is the use of lots of data divided over various leagues. So, researchers who focussed more on one league were able to obtain better results as their model could probably be trained more. Next to this, another limitation of this research is that it could not find any solution to the problem of predicting draws.

References

Adam, A. (2016) Generalised linear model for football matches prediction. *KULeuven, Departement of Computer Science.*

Baboota, R., & Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for Premier League. *International Journal of Forecasting.*

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *InProceedings of the 22nd ACM SIGKDD international conference on knowledgediscovery and data mining, KDD '16(pp. 785–794)*

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, *36*, 322-339.

Gomes, J., Portela, F., & Santos, M. F. (2015). Decision Support System for predicting Football Game result. In *Computers-19th International Conference on Circuits, Systems, Communications and Computers-Intelligent Systems and Applications Special Sessions. Series* (Vol. 32, pp. 348-353).

Harris, N. (2010). *The Premier League: Where the money went in 2009-10*. Retrieved from https://www.sportingintelligence.com/2010/05/16/the-premier-league-where-the-money-went-in-2009-10/

Kampakis, S., Thomas, W. Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches. *University College London*. Pages 19

Lock, S. (2010). *Sports Betting - Statistics & Facts*. Retrieved from https://www.statista.com/ topics/1740/sports-betting/

Min, B., Kim, J., Choe, C., Eom, H., McKay, R.B. "A compound framework for sports results prediction: A football case study," Knowledge-Based Systems, vol. 21, no. 7, pp. 551 – 562, 2008.

Moore, J. (2020). *Liverpool to get record £175m boost as Premier League clubs decide prize money cuts will not affect 2019/20 season.* Retrieved from https://talksport.com/football/745035/liverpool-premier-league-prize-money-2019-20-season/

New York Times. (2017). How Arsenal and Arsène Wenger Bought Into Analytics. Retrieved from *https://www.nytimes.com/2017/02/03/sports/soccer/arsenal-arsene-wenger-analytics.html*

Research and markets (2020). Researchandmarkets.com. *Sports Betting Market - Forecasts from 2020 to 2025, August edition.* https://www.researchandmarkets.com/reports/5138739/sports-betting-market-forecasts-from-2020

Tavakol, M., Zafartavanaelmi, H. and Brefeld, U. Feature Extraction and Aggregation for Predicting the Euro 2016. *Leuphana University of Luneburg*, 2016. Pages 19

Tax, N., Joustra, Y. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. *Transactions on Knowledge and Data Engineering*, 2015. pages 19

Transfermarkt. (2020). Transfermarkt.com. *Transfer Records.* Retrieved from: https://www.transfermarkt.com/statistik/transferrekorde

Transfermarkt. (2020). Transfermarkt.com. *League standings for the Spanish, English, German, Italian and French leagues.* Retrieved from http://www.transfermarkt.com

Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting Soccer Match Results in the Premier League *(Doctoral dissertation, Doctoral dissertation, Ph. D. dissertation, Stanford).*

Appendix

Table A1. *An overview of the reviewed studies regarding which divisions and seasons are used, which classifiers are used, and what prediction accuracy is obtained.*

Literature	Country	Classifiers	Accuracy
Ulmer et al. (2013)	England	Random Forest, SVM	0.50, 0.51
Constantinou et al. (2012)	England	Bayesian Networks	0.51
Gomes et al. (2015)	England	SVM	0.51
Tax et al. (2015)	Netherlands	Random Forest	0.55
Baboota and Kaur (2018)	England	Random Forest, Gradient Boost	0.57, ….
Tavakol et al. (2016)	International matches	Linear model	-
Min et al. (2008)	International matches	Bayesian Net	

Table A2. *An overview of the reviewed studies regarding which features were used.*

Study	Features that are used similar to this research
Ulmer et al. (2013)	Team Ranking + Difference in goals + Form in last matches (different streak)
Constantinou et al. (2012)	Team ranking in previous season + Form in last five matches
Gomes et al. (2015)	Goals and Shots average + Form in last five matches
Tax and Joustra (2015)	Average over Results & Goals + Form in last five matches
Baboota and Kaur (2018)	Team Attributes + Various match statistics + Form in last matches
Min et al. (2008)	Defensive Grade, Offensive grade, positioning grade
Tavakol et al. (2016)	Player-statistics
Adam (2016)	Player-statistics

Table A3: Football clubs used in French league with league rankings

Team	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15
Olympique Lyonnais	1	3	2	3	4	3	5	2
Girondins de Bordeaux	2	1	6	7	5	7	7	6
Olympique de Marseille	3	2	1	2	10	2	6	4
AS Nancy-Lorraine	4	15	12	13	11	18		
AS Saint-Étienne	5	17	17	10	7	5	4	5
Stade Rennais FC	6	7	9	6	6	13	12	9
LOSC Lille	7	5	4	1	3	6	3	8
OGC Nice	8	9	15	17	13	4	17	11
Le Mans FC	9	16	18					
FC Lorient	10	10	7	11	17	8	8	16
SM Caen	11	18		15	18			13
AS Monaco	12	11	8	18			2	3
Valenciennes FC	13	12	10	12	12	11	19	
FC Sochaux-Montbéliard	14	14	16	5	14	15	18	
AJ Auxerre	15	8	3	9	20			
Paris Saint-Germain	16	6	13	4	2	1	1	1
RC Lens	17		11	19				20
FC Metz	20							19
AC Ajaccio					16	17	20	
AC Arles-Avignon				20				
Angers SCO								
Dijon FCO					19			
En Avant de Guingamp							16	10
ES Troyes AC						19		
Évian Thonon Gaillard FC					9	16	14	18
FC Nantes		19					13	14
GFC Ajaccio								
Grenoble Foot 38		13	20					
Le Havre AC		20						
Montpellier Hérault SC			5	14	1	9	15	7
SC Bastia						12	10	12
Stade Brestois 29				16	15	20		
Stade de Reims						14	11	15
Toulouse FC		4	14	8	8	10	9	17
US Boulogne Cote D'Opale			19					

Table A4: Football clubs used in German league with league rankings

Team	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15
FC Bayern Munich	1	2	1	3	2	1	1	1
Bayer 04 Leverkusen	7	9	4	2	5	3	4	4
FC Schalke 04	3	8	2	14	3	4	3	6
VfL Wolfsburg	5	1	8	15	8	11	5	2
DSC Arminia Bielefeld	15	18						
FC Energie Cottbus	14	16						
Karlsruher SC	11	17						
Eintracht Frankfurt	9	13	10	17		6	13	9
Borussia Mönchengladbach		15	12	16	4	8	6	3
Hannover 96	8	11	15	4	7	9	10	13
Hamburger SV	4	5	7	8	15	7	16	16
Borussia Dortmund	13	6	5	1	1	2	2	7
VfB Stuttgart	6	3	6	12	6	12	15	14
SV Werder Bremen	2	10	3	13	9	14	12	10
Hertha BSC Berlin	19	4	18	1	16	1	11	15
TSG 1899 Hoffenheim		7	11	11	11	16	9	8
1. FC Köln		12	13	10	17			12
VfL Bochum	12	14	17					
1. FC Nürnberg	16		16	6	10	10	17	
1. FSV Mainz 05			9	5	13	13	7	11
SC Freiburg			14	9	12	5	14	17
1. FC Kaiserslautern				7	18			
FC St. Pauli				18				
FC Augsburg					14	15	8	5
SpVgg Greuther Fürth						18		
Fortuna Düsseldorf						17		
Eintracht Braunschweig							18	

SC Paderborn 07								18
SV Darmstadt 98								
FC Ingolstadt 04								

Table A5: Football clubs used in Italian league with league rankings

Team	**2007/08**	**2008/09**	**2009/10**	**2010/11**	**2011/12**	**2012/13**	**2013/14**	**2014/15**
Atalanta	9	11	18		12	15	11	17
Bari			10	20				
Bologna		17	17	16	9	13	19	
Brescia				19				
Cagliari	14	9	16	14	15	11	15	18
Carpi								
Catania	17	15	13	13	11	8	18	
Cesena				15	20			19
Chievo Verona		16	14	11	10	12	16	14
Empoli	18							15
Fiorentina	4	4	11	9	13	4	4	4
Frosinone								
Genoa	10	5	9	10	17	17	14	6
Hellas Verona							10	13
Inter	1	1	1	2	6	9	5	8
Juventus	3	2	7	7	1	1	1	1
Lazio	12	10	12	5	4	7	9	3
Lecce		20		17	18			
Livorno	20		20				20	
Milan	5	3	3	1	2	3	8	10
Napoli	8	12	6	3	5	2	3	5
Novara					19			
Palermo	11	8	5	8	16	18		11
Parma	19		8	12	8	10	6	20
Pescara						20		
Reggio Calabria	16	19						
Roma	2	6	2	6	7	6	2	2
Sampdoria	6	13	4	18		14	12	7
Sassuolo							17	12
Siena	13	14	19		14	19		
Torino	15	18				16	7	9
Udinese	7	7	16	4	3	5	13	16

Table A6: Football clubs used in English league with league rankings

Team	**2007/08**	**2008/09**	**2009/10**	**2010/11**	**2011/12**	**2012/13**	**2013/14**	**2014/15**
Arsenal	3	4	3	4	3	4	4	3
Aston Villa	6	6	6	9	16	15	15	17
Birmingham	19		9	18				
Blackburn	7	15	10	15	19			
Blackpool				19				
Bolton	16	13	14	14	18			
Burnley			18					19
Cardiff							20	
Chelsea	2	3	1	2	6	3	3	1
Crystal Palace							11	10
Derby	20							
Everton	5	5	8	7	7	6	5	11
Fulham	17	7	12	8	9	12	19	
Hull		17	19				16	18
Leicester								14
Liverpool	4	2	7	6	8	7	2	6
Man City	9	10	5	3	1	2	1	2
Man United	1	1	2	1	2	1	7	4
Middlesboro	13	19						
Middlesbrough	13	19						
Newcastle	12	18		12	5	16	10	15
Norwich					12	11	18	
Portsmouth	8	14	20					
QPR					17	20		
Reading	18					19		20

Southampton						14	8	7
Stoke		12	11	13	14	13	9	9
Sunderland	15	16	13	10	13	17	14	16
Swansea					11	9	12	8
Tottenham	11	8	4	5	4	5	6	5
Watford								
West Brom		20		11	10	8	17	13
West Ham	10	9	17	20		10	13	12
Wigan	14	11	16	16	15	18		
Wolves			15	17	20			

Table A7: Football clubs used in Spanish league with league rankings

Team	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15
Athletic Club de Bilbao	11	13	8	6	10	12	4	7
Atlético Madrid	5	4	9	5	5	3	1	3
CA Osasuna	17	14	12	9	7	16	18	
CD Numancia		19						
CD Tenerife			19					
Córdoba CF								20
Elche CF							16	13
FC Barcelona	3	1	1	1	2	1	2	1
Getafe CF	13	17	6	16	13	10	14	14
Granada CF					17	15	15	18
Hércules Club de Fútbol				19				
Levante UD	20			13	6	11	10	15
Málaga CF		8	17	11	4	6	11	9
Racing Santander	6	11	16	12	20			
Rayo Vallecano					15	8	12	11
RC Celta de Vigo						17	9	8
RC Deportivo de La Coruña	8	7	10	18		19		17
RC Recreativo	16	20						
RCD Espanyol	12	10	11	8	14	13	13	10
RCD Mallorca	7	9	5	17	8	18		
Real Betis Balompié	14	18			12	7	20	
Real Madrid CF	1	2	2	2	1	2	3	2
Real Sociedad				15	11	4	7	12
Real Sporting de Gijón		16	15	10	19			
Real Valladolid	15	15	18			14	19	
Real Zaragoza	18		14	14	16	20		
SD Eibar								16
Sevilla FC	4	3	4	7	9	9	5	5
UD Almería	9	12	13	20			17	19
UD Las Palmas								
Valencia CF	10	6	3	3	3	5	8	4
Villarreal CF	2	5	7	4	18		6	6
Xerez Club Deportivo			20					

Table A8: Python packages used in this research

Code in Anaconda Navigator
import sqlite3
import numpy as np
import pandas as pd
from datetime import datetime as dt
import xml.etree.ElementTree as ET
from sklearn.model_selection import TimeSeriesSplit, train_test_split
from sklearn.linear_model import LogisticRegression, RandomForestClassifier
import xgboost as xgb
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report, accuracy_score
from sklearn.calibration import CalibratedClassifierCV

```
from sklearn import model_selection
from sklearn.model_selection import train_test_split
from sklearn.metrics import make_scorer
from time import time
from sklearn.decomposition import PCA, FastICA
import warnings
from functools import reduce
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import TimeSeriesSplit, train_test_split
from sklearn.metrics import roc_auc_score
import itertools
from IPython.display import display
```