PREDICTING FOOTBALL MATCH OUTCOMES USING MACHINE LEARNING ALGORITHMS

Menno Heijboer
SNR: 2001009
M.r.heijboer@tilburguniversity.edu

Word count: 8769

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. Sharon Ong
Dr. Federico Zamberlan

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
May 20, 2022

Preface

Dear reader,

I hereby present to you my thesis, in which machine learning algorithms were utilized to classify the outcome of football matches. I would like to thank my supervisor dr. Sharon Ong for her feedback and pragmatic approach to challenges throughout this process. Finally, I would like to thank my parents for their continuous support.

Abstract

Accurately predicting football matches can have significant economic consequences, with the value of the football betting industry at an all-time high. From a scientific standpoint, more research is needed on whether matches ending in a tie can be predicted. This outcome was primarily misclassified or ignored in previous research. This study uses Machine Learning algorithms and resampling techniques on a dataset containing 6237 matches, collected over five competitions between 2008 and 2016. The main target is to explore the extent to which it is possible to predict the outcome of football matches, which is done in a multi-class manner. In addition, a comparative approach is taken to reflect on the performance of algorithms, features, operationalizations, and resampling methods. The maximum extent to which matches could be predicted was 53.7% accurate and obtained using a Random Forest, which outperformed bookmakers' predictions by 0.7%. The best results on ties are obtained using a Gradient Booster combined with ClusterCentroids undersampling (F1 = 0.37). Ratings from FIFA were the most informative feature during predictions, outperforming related metrics such as the ELO rating. Lastly, features calculated as differential scores generally outperformed separate features for home- and away teams. A major limitation of this study is that algorithms are capable of identifying the stronger team, but do not seem to include scenarios in which the stronger team ends up losing. For future studies, clustering algorithms might be useful to explore patterns within matches won by the underdog, which could add complexity to the decision-making of classifiers.

Keywords: Football, Multi-class prediction, Supervised Machine Learning, Random Forest, Gradient Boosting, Support Vector Machine, Logistic Regression, ELO, FIFA

1. Ethics Statement

Work on this thesis did not involve collecting data from human participants or animals. The database used during this study has been made publicly accessible. The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis.

2 Introduction

2.1 Background and Project Definition

Data experts are becoming football's best signings. These words stem from a BBC article on March 5^{th,} 2021, describing the increased use of data and the value of skilled data analysts in football's current landscape (Harper, 2021). As the gap in financial resources between clubs increases (Rohde & Breuer, 2016), not all teams have the luxury to make valid use of available data, which puts smaller clubs on the back foot. A similar trend is noticeable in the betting industry, where large bookmakers have had the knowledge and resources to make good use of data for some time (Graham & Scott, 2008), whereas individuals placing bets have limited access to resources and less insight into which data to monitor. To make data-driven decision-making more accessible and less time-consuming, the objective of this thesis is to find suitable machine learning approaches to predict the outcome of football matches and determine which data streams are most important to monitor.

2.2 Motivation

Addressing the current inequality in options for data usage is worthwhile on multiple levels. It can create a more level playing field between organizations with different budgets, help individuals make an informed decision when placing bets, and can contribute to our collective understanding of what factors influences the outcome of football matches. The ability to predict football matches, or at least to a degree that it is competitive to bookmakers, can have a significant economic impact. In 2013, the value of the football betting industry was already estimated to be around 700 billion up to 1 trillion dollars (Keogh & Rose, 2013). Besides economic opportunities, from a scientific point of view, it will be worthwhile to combine new data science techniques with the increasingly diverse nature of football-related data (Constantinou & Fenton, 2017) and address some of the shortcomings in existing literature. As of now, the number of studies predicting football matches is still low.

2.3 Research Questions

This study used supervised machine learning (ML) techniques to predict the outcome of football matches in a multi-class manner, with outcomes being either a home team win, a tie/draw, or away team win. A comparative approach was taken to optimize features and compare different methodologies, algorithms, and resampling strategies. Figure 1 gives an overview of the workflow throughout this thesis. The main target was to explore the research question given below.

RQ: To what extent can the outcome of football matches be predicted using machine learning algorithms

As one of the most used metrics to quantify a team's strength, several implications accompany the use of ELO scores in club football. This study will compare the predictive capabilities of the original ELO rating with alternative approaches, and reflect on its usefulness in association football.

SQ1: Can the predictive capabilities of the traditional ELO rating as a reflection of a team's strength be improved using personalized starting ratings?

In related studies, various approaches were taken to operationalize match events. As there is no clear consensus on what method leads to the optimal outcome, these approaches were compared to a method introduced in this thesis.

SQ2: What method of operationalization for match event data yields the best outcome?

Since insufficient focus has been placed on individual feature importance during predictions, it is easy to get lost in irrelevant data streams. More insight into what features are worthwhile to monitor can help prioritize certain data types.

SQ3: What features are most important during classification?

Four machine learning classifiers are compared in order to find the best-equipped algorithm for match outcome predictions. These algorithms will be tested under varying circumstances.

SQ4: What classifier is most suitable for the prediction of football matches?

To improve the predictions of lesser occurring outcomes, three resampling methods are applied. The objective of these efforts is to force classifiers to place equal emphasis on all possible outcomes and reduce bias towards the majority class. Especially draw outcomes were proven difficult to classify and were largely ignored in previous studies.

SQ5: Can the predictions of a draw as a match outcome be improved by use of balancing methods to mitigate the class imbalance in the match outcome distribution

3. Related Work

3.1 ELO Rating

The ELO rating is a measurement originating in chess (Elo, 1978), and later modified to fit the football game (Buchdahl, 2003; Hvattum & Arntzen, 2010). Most recently, Herbinet (2018) explored an alternative approach to ELO scores, by including information on match events during calculation. Subsequently, they created personalized initiation values for teams at the beginning of each season by dividing the average goals scored by goals conceded during the previous season. Besides their experiments with ELO calculations, Herbinet (2018) compared Naïve Bayes, Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), Neural Network, and K-Nearest-Neighbor models. A linear SVM was the best performer with 51% accuracy. A limitation in this study is the use of data stemming from previous seasons to initialize ELO ratings. Player squads are often subject to significant personal changes in-between seasons (FIFA, 2021), which makes past data somewhat unreliable. However, a shared initiation value among teams creates more issues, given that ELO ratings require around 20 to 30 matches to converge towards a team's actual strength (Krifa et al., 2021; World Football Elo Ratings, 2022), while regular seasons comprise less than 40 matches. To address both issues, a solution could be a more player-based approach with the use of individual player ratings from EA Sports FIFA game. The usefulness of FIFA ratings for match classification has been demonstrated in previous studies (Prasetio, 2016; Matano et al., 2018; Chen, 2019) and can initiate ELO ratings based on the quality of individual players.

3.2 Match Events

Another type of variable that has been used in football-related studies is match events. A match event can be seen as a specific type of event occurring in a football match, such as shots, corners, and fouls. They have been applied in previous studies to distinguish between player roles (Aalbers & Van Haaren, 2018), explore gender differences (Garnica-Caparrós & Memmert, 2021), quantify a team's playing style (Riezebos et al., 2021), and predict outcomes (Eryarsoy & Delen, 2019; Baboota & Kaur, 2018). Both studies that used events for outcome prediction operationalized these events differently. Baboota and Kaur (2018) used these events as the absolute occurrence during the previous game, whereas Eryarsoy and Delen (2019) used the weighted average occurrence of events over all previous matches. As for now, it is not yet clear which method yields better performance.

3.3 Football Outcome Prediction

Techniques used in older papers were not comparable, hence only recent studies using ML will be reflected upon. One of these studies was done by Tax and Joustra (2015), in which a wide array of variables such as *bookmakers' odds*, *performance ratios*, *streaks*, *home advantage*, *promotion status*, and *absence of key players* were used. Random Forest (RF) and Logit Boosting algorithms were trained on 65 features, with both algorithms showing similar performance around 55% accuracy. A noteworthy takeaway was that models using only bookmakers' odds were outperformed by models including all features. Another study to include a wide array of variables is Baboota and Kaur (2018), with variables such as *from, attacking- and defensive ratings, streaks, match events,* and *goal difference*. Ensemble methods such as the Gradient Booster (GB) and RF were best equipped to fit the high-dimensional dataset, with accuracy scores around 56.5%. Furthermore, they concluded that using differential features superior compared to separate features for home and away teams, given the better univariate distribution and reduced number of dimensions. Most recently, Eryarsoy and Delen (2019) adopted a multi-variate approach for predictions while also reflecting on the individual feature performance. They included 53 variables describing various *team characteristics* and *match events*. Their goal was to compare singular methods to ensemble methods. As was the case in similar studies, ensemble methods such as the RF and GB once again proved superior to singular approaches, with both ensemble models achieving 74% accuracy. This performance raises suspicion when compared to related studies, especially when looking at the recall (70.8%) and precision (66.6%) in predicting matches that end in a tie. These inflated results might be caused by noise introduction, missing value imputation, or incorrect resampling. Eryarsoy and Delen (2019) was the first study to address the poor performance on draw outcomes by applying SMOTE (Chawla et al., 2002). It is crucial to only use these synthetically created data points during training since algorithms often recognize these points, leading to inflated performance during testing (Muralidhar, 2021). As for individual feature performance, they found that *percentage of possible points* earned, *current league standing*, *formation consistenc*y, *promotion status,* and *age* had high predictive capabilities for football match outcome.

3.4 Addressing the Class Imbalance in Match Outcomes

Since most algorithms try to achieve maximum classification accuracy, too much focus is being placed on data points belonging to the majority class, as this approach often leads to the optimal accuracy in an imbalanced dataset (Reza & Ma, 2018). This trend has been noticeable in studies predicting football matches in a multi-class manner, which all display a bias towards home team victories and (almost) entirely ignored ties. There are several ways to go about imbalances in the data, like increasing the amount of minority data points. This approach was chosen by Eryarsoy and Delen (2019) in the form of SMOTE (Chawla et al., 2002), generating synthetic copies of matches ending in a draw or away team win. Contrary to oversampling, another approach is to leave out data points belonging to the majority class. Zhang et al. (2010) compared undersampling methods and concluded that clustering approaches outperformed random selection of majority data points to drop. Both approaches however come with limitations. Farid et al. (2016) found that undersampling often leads to significant loss of information, whereas oversampling techniques still lead to increased overfitting risk (Last et al., 2017). A third approach not reliant on the creation or deletion of data is the addition of a penalty for misclassifying minority data points (Johnson & Khoshgoftaar, 2019), forcing the algorithms to place equal emphasis on all classes. Since only one study so far has addressed the problems caused by imbalance, it is currently unknown what approach is best suited for football-related studies.

3.5 Limitations and Contribution

This study will be an opportunity to address some of the shortcomings in existing literature, while simultaneously comparing their methodology. Many studies suffered from a lack of generalization potential, had problems predicting the lesser occurring match outcomes, and underreported feature importance on an individual level. By using a dataset consisting of matches from a variety of leagues, applying various balancing techniques, and using algorithms with the ability to display how their predictions came about, this study can make a contribution to existing literature and serve as a benchmark for future research.

4. Methodology

4.1 Dataset

The dataset used for this study is freely accessible on Kaggle and is called the '*European Soccer Database'* (Kaggle, 2016). It comes in SQL format, has a size of 313.09 MB, and is last updated in 2016. The original data contains information from over 25.000 football games, collected in eleven competitions, spanning over 12 years. It holds five information sources, described in table 1. Match events were obtained from a supplemental database called the *'European Soccer Database Supplementary'* (Kaggle, 2017). In the original database, match events were encoded in XML format, which was transformed into CSV files in the supplementary database. For the current study, 6237 matches fitted the criteria for inclusion based on the presence of match events. These matches were collected in six competitions and comprise information from seventeen seasons. The exact number of games per league can be found in appendix 1.1.

Table 1

Sources of information comprising the original database

Data type	Description
Match events	Contains events occurring during football matches, such as goals, shots on- and off-target, crosses, possession, fouls, and bookings. Also contains additional information about the specific types of events
Match information	Containing participating teams, starting player line-ups, formations, and additional information such as the league, stage, and season in which the match took place.
Betting odds	Odds by several of the largest bookmakers for all possible match outcomes per individual match.

Team- and player ratings	Data collected from the official EA Sports FIFA game. This dataset was updated regularly since FIFA updated its ratings based on the real-life performance of players and teams.
Keys	Each league, team, player, match, and season's unique identifier key

4.2 Software and Libraries

DB browser was used to open and transform the database. Calculations were made in Python version 3.9.0, by use of Anaconda and Jupyter Notebook 6.3.0. To extract information from a variety of tables into one CSV file, the libraries *Pandas* and *Numpy* were utilized. The application of algorithms and subsequent evaluation of algorithms were done with *Sklearn.* Visualizations were made with *Pyplot, Seaborn,* and *Yellowbrick*. Finally, all resampling efforts were done with the *Imblearn* library.

4.3 Workflow

Since this study contains various steps before being able to reflect on the research question and not all sub-questions are answered at the same moment, an overview of the workflow throughout this thesis is given in figure 1 on the next page.

4.4 Data Extraction

The SQL database was loaded into *DB Browser*, from which the separate tables were extracted as CSV files. The transformation from multiple SQL tables in CSV format into one workable CSV file was done manually with *Numpy* and *Pandas* libraries. The table containing match information was used as the main data frame in which the extracted information from other tables was stored. The unique identifier keys made it possible to link all corresponding information.

4.5 Pre-processing

After the selection of matches based on the presence of match events, the final dataset was relatively clean when it came to missing values. 1738 matches had no information on the distribution of *possession* between teams. Those values were supplemented with the KNN missing value imputation function from the *Sklearn* library, which operates by imputing missing values with the mean score of the *n* neighbors (Obadia, 2017). In addition, a few player ratings were imputed by the average team rating. Average odds from bookmakers were calculated solely on the available odds, so there was no need to impute missing values. As the final step of preprocessing, data were divided into a training and a test set with the *Sklearn* train-test split function. 70% of data was selected to train algorithms, whereas the other 30% was set aside for evaluation.

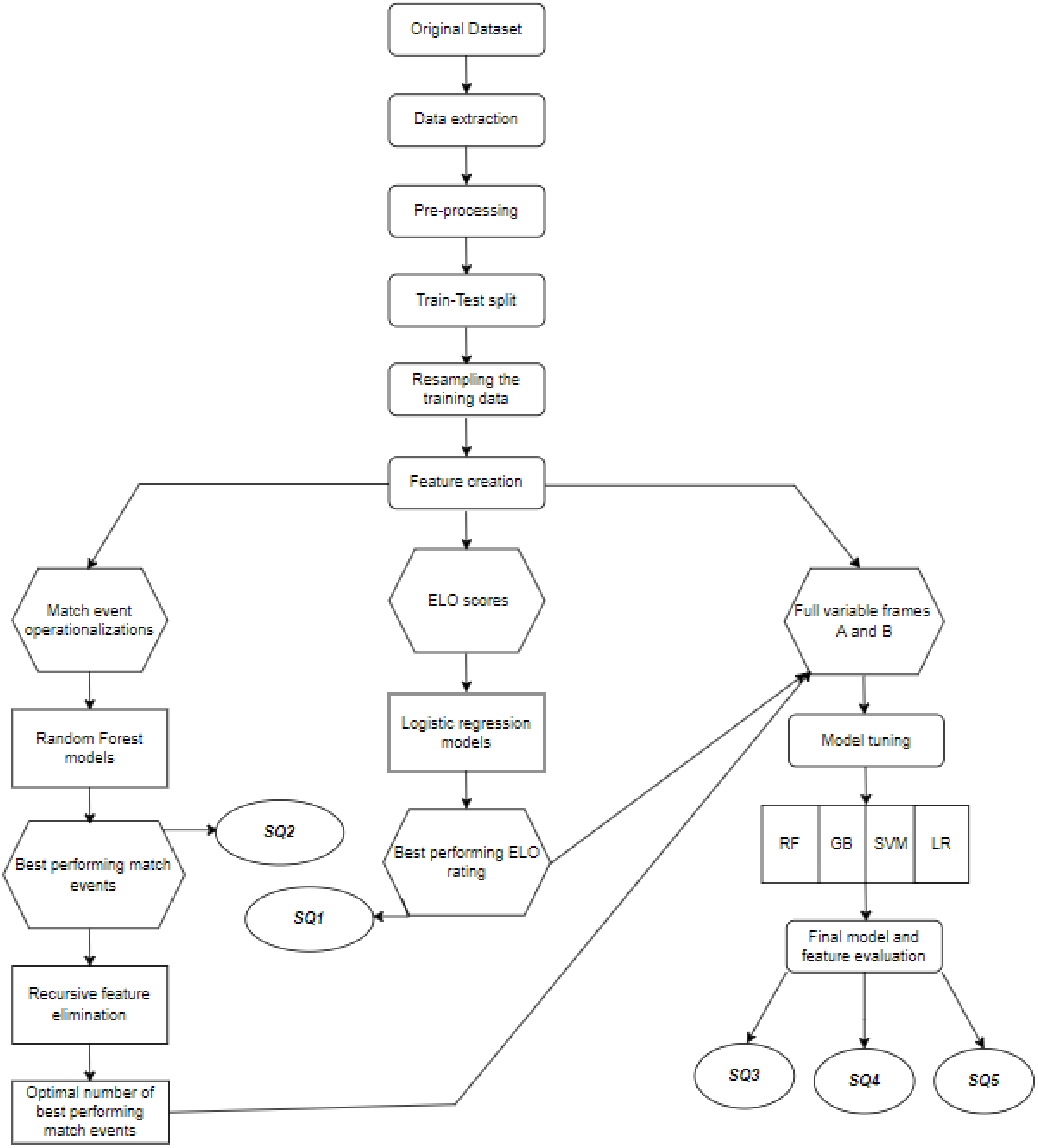

Original Dataset
Data extraction
Pre-processing
Train-Test split
Resampling the training data
Feature creation
Match event operationalizations
ELO scores
Full variable frames A and B
Random Forest models
Logistic regression models
Model tuning
Best performing match events
SQ2
Best performing ELO rating
RF
GB
SVM
LR
SQ1
Final model and feature evaluation
Recursive feature elimination
Optimal number of best performing match events
SQ3
SQ4
SQ5

Figure 1. Workflow throughout this study

4.6 Resampling

Training data was modified using over-and undersampling techniques. As far as known, only one related paper has put effort into dealing with the unbalanced distribution of outcomes. To mitigate this imbalance (displayed in figure 2), two resampling techniques were proposed. The first method is called the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al., 2012), in which new data points are synthetically created based on scores of neighboring data points, and are therefore not exact replicas of original data. This reduces the risk of overfitting while simultaneously making it easier for algorithms to fit a decision boundary (Seo & Kim, 2018). The second method is an undersampling approach called ClusterCentroids (CC), in which a new dataset is generated based on the cluster centroid of a K-means algorithm. One advantage that comes with CC is that it only transforms the majority class, while the minority classes preserve themselves (Yagci, 2021). A more in-depth description of the workings of the K-means algorithm can be found in an article by Dabbura (2021). Even though both techniques do not directly copy existing data, they still bring about an increased risk of either overfitting or underfitting (Farid et al., 2016; Last et al., 2017). Therefore, a third approach was selected that did not alter the original data, but rather changes the workings of algorithms by adding equal class weights. Algorithms in *sklearn* allow for class weights to be set, forcing them to put equal emphasis on all classes.

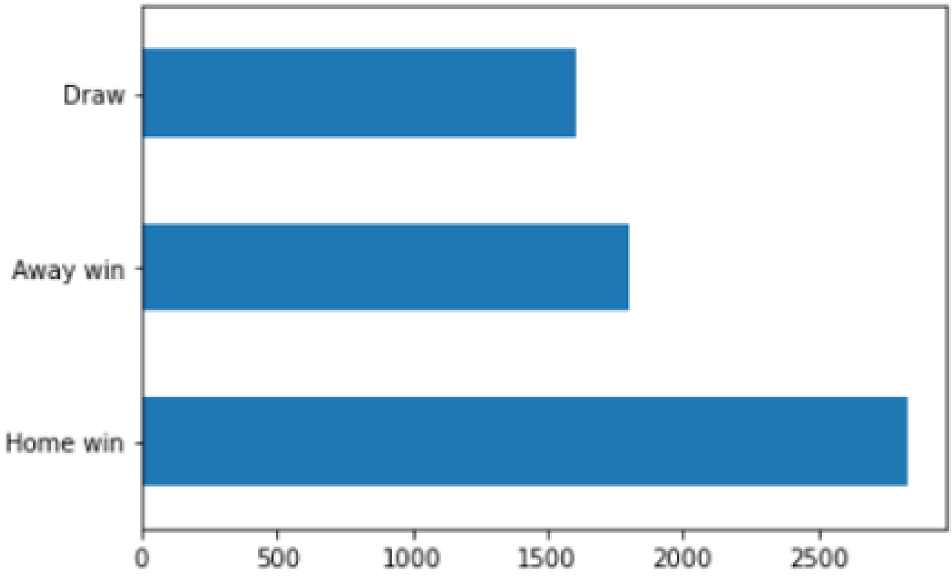


Figure 2. Distribution of match outcomes

4.7 Feature Creation

During the process of feature creation, much effort was put into making sure that only information was used that was available before a match took place, in order to avoid introducing irrelevant noise. Besides promotion status, no information from previous seasons was passed on.

4.7.1 ELO Rating

For the first sub-question, the original ELO rating was compared to alternate approaches. The formula for the ELO rating is given in Eq. (1) below:

$$Rating_{new} = Rating_{old} + (K + K \times GD) \times (Result - Expected) \quad (1)$$

The old rating reflects a team's rating before a match and is initialized at a set value of 1500 (Kovalchik, 2020; Aldous, 2017; Bisberg & Rivera, 2019). This value serves as the average rating of all teams within a competition. K determines the step size when updating ratings after a match. This value was initially set at 50 since this was found the optimal base K in a cross-validation study of ELO parameters (Bisberg & Rivera, 2019), but different values were tried later on. GD stands for the factor by which K is multiplied, based on the score difference. Score differences equal to one, two, and three result in a GD of 0, ½, and ¾ respectively. For matches with score differences equal to or larger than four, GD is equal to ¾ + (N – 3)/8, in which N reflects the actual score difference. The result variable is determined by the match outcome, where a win is awarded one point, half a point is given for a tie, and 0 points for a loss. The expected variable requires a separate formula, which is given in Eq. (2) below. In this formula, the rating difference is calculated by subtracting the away team's pre-match ELO rating from that of the home team. To take home-team advantage (HTA) into account, the customary approach is to add an additional 100 points to the home-team rating (World Football Elo Ratings, 2022). Similar to K, the HTA was tried with different values to find its optimal setting.

$$Expected = \frac{1}{10^{(-rating\ difference\ \div 400)} + 1} \quad (2)$$

The original ELO rating was compared to an alternate approach, which throughout this thesis is called the *ELO proxy* rating. This proxy rating was calculated in an identical fashion, with the exception that FIFA ratings were used to create a personalized initiation value at the start of a season. The average overall rating of a team's starting line-up during the first match (M = 75.146, SD = 3.947) was multiplied by 20 to match the standard ELO scale. The applicability of FIFA ratings has been demonstrated in previous studies (Chen, 2019; Arntzen & Hvattum, 2021; Matano et al., 2018), and was tried as a method to tackle convergence issues that accompany the ELO rating in club football without the use of past data. As mentioned, different values were tried for K (10, 25, 50, 75, 100) and the home team advantage (0, 50, 100, 150, 200). The combinations resulting in the highest average Spearman's rank correlation coefficient between ratings and match outcomes were compared to their counterparts with default settings K = 50 and HTA = 100.

4.7.2 Match Events

For the second sub-question, the aim was to explore which approach was best equipped to operationalize match event data. The original dataset consisted of seven types of match events (see appendix 1.2), which were expressed as offensive (occurrence of an event by a team) and defensive (occurrence of an event against a team) variables. Since variables had to be calculated for both home- and away teams as well, the total number of match events ended up being 28. These events were either calculated as weighted average occurrence over the entire season (Eryarsoy & Delen, 2019), the absolute occurrence during the last match (Baboota & Kaur, 2018), or the weighted average occurrence over the last 5 matches. This final approach is unique to the current study and introduces a momentum factor into the average, in which recent matches influence the average to a greater extent compared to seasonal averages This was desired given the effects brought about by factors such as *form* and *win/losing streak* (Heuer & Rubner, 2009; Goddard, 2006). During the first 5 stages of a season, averages were calculated over the number of matches available.

4.7.3 Full Variables Frames

The aim of the fourth sub-question was to predict the outcome of football matches using a wide array of variables, in order to identify the best-performing features during classification. The variables used were largely inspired by related studies, supplemented by several features unique to the current research. Some of these unique features are *aggression*, *average weight*, *average BMI*, the *ratio of the 4 most drafted players* in the starting line-up, the number of *changes in the starting line-up* compared to the previous game, and several *shooting metrics* (MacDonald, 2012). Information on all variables created, the method of calculation, and related papers to include similar variables can be found in appendix 1.7. All numeric variables were either calculated separately for home- and away teams (model A), or as a differential score (model B). Lastly, all variables used in the final models were scaled between 0 and 1 with the Min-Max normalization formula, given in Eq. 3. Since the differential scores were a product of subtraction, those features will range between -1 and 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

4.8 Feature Selection

The classification performance of different ELO calculations (*ELO original, ELO proxy, optimized ELO original,* and *optimized ELO proxy*) was measured using LR models. This method was chosen since regression models are understandable, fast, and easy to create (Swaminathan, 2018), which sufficed given the inclusion of only two variables per model. The best performing rating in terms of accuracy was adopted into the final models.

The classification accuracy of three types of match event operationalizations (*weighted average over season*, *weighted average over 5 games*, *absolute occurrence previous game*) was assessed using RF models. Since several match events often occur together due to the nature and rules of football, RF models were deemed fitting given that they work by selecting different features for each iteration, avoiding overfitting and robust to collinearity (Matsuki et al., 2016). A collinearity plot between events can be found in appendix 2.1. After selecting the best performing operationalization, Recursive Feature Elimination (RFE) was applied using *Yellowbrick's* RFECV model, to determine the optimal number of events to retain. This model starts training on a full feature set and recursively drops features to prune insignificant variables (Bex, 2021). The specified number of most informative events were included in the final models. Importance was determined by the average decrease in impurity per event.

For the final models, all variables had to be created manually and can be found in appendix 1.7. Even though variables such as the *ELO rating*, *form*, *percentage of points obtained*, *current ranking*, and *goal difference* showed high collinearity, the aim of this multi-variate approach was to explore which variables contributed the most when predicting football matches. Hence, no variables were left out, but this collinearity was dealt with by selecting algorithms that were more robust in dealing with this issue. Collinearity plots between home team variables, away team variables, and differential variables can be found in appendices 2.2, 2.3, and 2.4.

4.9 Algorithms

As for the selection of algorithms, all techniques had to meet two requirements: they must have the ability to display their inner workings and must be suited to deal with high-dimensional data. Four ML techniques were selected: a Random Forest, Gradient Booster, Support Vector Machine, and Logistic Regression. A quality these algorithms share is the option to display how classifications came about, compared to a more black-box approach in Neural Networks (Zhang et al., 2018). This clarity in decision-making was relevant due to the exploratory nature of this study. More detailed reasoning behind algorithm selections is given below. For each algorithm, the optimal set of hyperparameters was sought out with a 5-fold cross-validation. A Grid containing all values was combined with the GridSearchCV function from *Sklearn*. The exact values used during tuning and optimal values can be found in tables 2 and 3. A short description of hyperparameter functions can be found in appendix 1.3.

4.9.1 Random Forest

This algorithm is an ensemble method, in which the results of individual decision trees are combined to make final predictions (Horning, 2010). Besides the fact that ensemble models often yield better performance than individual models (Dahinden & Ethz, 2011), RF models have the advantage that they are less likely to overfit training data compared to an individual decision tree. In addition, RFs are good with high dimensional data and offer quick prediction- and training speed (Kho, 2018).

4.9.2 Gradient Boosting

A Gradient Boosting algorithm is one of the more powerful algorithms in ML and is often used to minimize the amount of bias error within models (Tarbani, 2021). It works by building simple models in sequential order, with each new model trying to minimize the error left over by the previous one. As for most ensemble methods, the GB algorithm is more flexible and less reliant on the type of input data compared to singular methods (Tarbani, 2021).

4.9.3 Support Vector Machine

An SVM can be applied to regression- and classification problems. It works by finding the optimal hyperplane in an N-dimensional space, which maximizes the distance between data points belonging to different classes (Gandhi, 2018). The value of N is determined by the number of variables within a model. A major advantage that comes with an SVM is that they are well-suited for multi-variate input, given their ability to work in high-dimensional spaces. Another advantage is the kernel hyperparameter, making SVMs useful for linear and non-linear problems (Herbinet, 2018).

4.9.4 Logistic Regression

This algorithm is a relatively simple classifier that applies a logistic function to model the relationship between a discrete dependent variable and one or more independent variables (Hilbe, 2009). Several advantages that accompany this method are fast execution speed, easy interpretation, and the ability to distinguish between feature importance by looking at model coefficients (Raj, 2020).

Table 2

Values tried for model A and optimal hyperparameter values

Classifier	Parameters used in GridSearch	Best parameters
Random Forest	Max_depth: 5, 10, 15, 20, 25, 30	Max_depth: 10
	Max_features: 2, 4, 6, 8, 10, 12	Max_features: 12
	Min_samples_split: 2, 6, 10, 14	Min_samples_split: 14
	N_estimators: 50, 250, 500, 1000, 1500, 2000, 2500	N_estimators: 2000
Gradient Booster	Learning_rate : 0.001, 0.01, 0.1, 1, 10, 100	Learning_rate: 0.1
	Max_depth: 5, 10, 15, 20, 25, 30	Max_depth: 13
	N_estimators: 50, 250, 500, 1000, 1500, 2000, 2500	N_estimators: 2500
	Subsample: 0.5, 0.75, 1	Subsample: 1
SVM	C: 0.001, 0.01, 0.1, 1, 10, 100	C: 0.1
	Gamma: 0.001, 0.01, 0.1, 1, 10, 100	Gamma: 0.001
	Kernel: ‘rbf’, ‘linear’, ‘poly’, ‘sigmoid’	Kernel: ‘linear
LR	C: 0.001, 0.01, 0.1, 1, 10, 100	C: 1
	Solver: ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘saga’	Solver: ‘lbfgs’

Table 3

Values tried for model B and optimal hyperparameter values

Classifier	Parameters used in GridSearch	Best parameters
Random Forest	Max_depth: 5, 10, 15, 20, 25, 30	Max_depth: 10
	Max_features: 2, 4, 6, 8, 10, 12	Max_features: 6
	Min_samples_split: 2, 6, 10, 14	Min_samples_split: 14
	N_estimators: 50, 250, 500, 1000, 1500, 2000, 2500	N_estimators: 1000
Gradient Booster	Learning_rate : 0.001, 0.01, 0.1, 1, 10, 100	Learning_rate: 0.1
	Max_depth: 5, 10, 15, 20, 25, 30	Max_depth: 8
	N_estimators: 50, 250, 500, 1000, 1500, 2000, 2500	N_estimators: 2000
	Subsample: 0.5, 0.75, 1	Subsample:1
SVM	C: 0.001, 0.01, 0.1, 1, 10, 100	C: 100
	Gamma: 0.001, 0.01, 0.1, 1, 10, 100	Gamma: 0.001
	Kernel: ‘rbf’, ‘linear’, ‘poly’, ‘sigmoid’	Kernel: ‘linear’
LR	C: 0.001, 0.01, 0.1, 1, 10, 100	C: 0.1
	Solver: ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘saga’	Solver: ‘liblinear’

4.10 Performance Evaluation

All models were evaluated based on their predictions of test data. A visualization of predictions per class was given by confusion matrices. During the comparisons of *ELO ratings* and *match events*, the only metric of interest was the overall accuracy of models, which is calculated by the formula in Eq. (4). To reflect upon the final models, the performance of each model was expressed by their overall accuracy, supplemented by weighted precision. To take poor predictions on ties into account, weighted average precision was chosen over the micro average. Precision gives the ratio of correctly predicted instances of a specific class divided by the total number of times this class was predicted. The formula for calculation can be found in Eq. (5). An increased level of precision was preferred over recall, given that missed opportunities are less of an issue compared to misclassified outcomes. During the final stage of modeling, algorithms were trained on a re-sampled batch of matches to improve the emphasis on a draw as a viable match outcome. The performance on this outcome was evaluated based on precision, recall, and the F1 score. Recall can be seen as the detection ratio of draws, whereas the F1 reflects a harmonious mean between precision and recall. Formulas to calculate these metrics are given in Eq. (6) and Eq. (7) below.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{4}$$

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \tag{5}$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

To finalize the reflection on model performance, outcomes will be compared to related studies and two baselines. The first baseline is a majority classifier, which predicted all instances to belong to the most occurring class. The second baseline was created based on bookmaker predictions. The average odds per match outcome from a variety of bookmakers were used to create predictions, with the lowest average odds seen as the most likely outcome.

5. Results

5.1 Summary of Results

Several classifiers were utilized to predict the outcome of football matches. The best-performing model for each algorithm in terms of accuracy is displayed in table 4, which also includes a comparison to baselines. The most accurate model was a Random Forest with an accuracy of 53.7%, obtained with differential features (*model B*) and without resampling. This model turned out to be the only model able to outperform the predictions made by bookmakers, by a margin of 0.7%. Furthermore, all models outperformed the majority baseline. The only algorithm to benefit from resampling in terms of accuracy was the Gradient Booster, which was also the only algorithm to take ties into account without resampling (see tables 4 and 12).

Table 4

Best performing models per algorithm in terms of overall accuracy

Algorithm	Feature set	Resampling	Accuracy	Odds baseline	Majority Baseline
Random Forest	*Model B*	*None*	53.7%	+0.7%	+8.2%
Gradient Booster	*Model B*	*SMOTE*	51.8%	-1.2%	+6.3%
Support Vector Machine	*Model B*	*None*	53.0%	-0.0%	+7.5%
Logistic Regression	*Model B*	*None*	53.0%	-0.0%	+7.5%

To gain more insight into predictions, confusion matrices are displayed in figure 3, which show the performance over separate classes (home win = 0, draw = 1, away win = 2) in the most accurate models. As was the case in related studies, all algorithms seemed to have a bias towards home team victories, whereas draws get largely ignored. An increased focus on a draw however rarely seems to increase the overall model performance. Moreover, the application of both a linear SVM and LR model seemed redundant looking at the similarity in predictions over all classes.

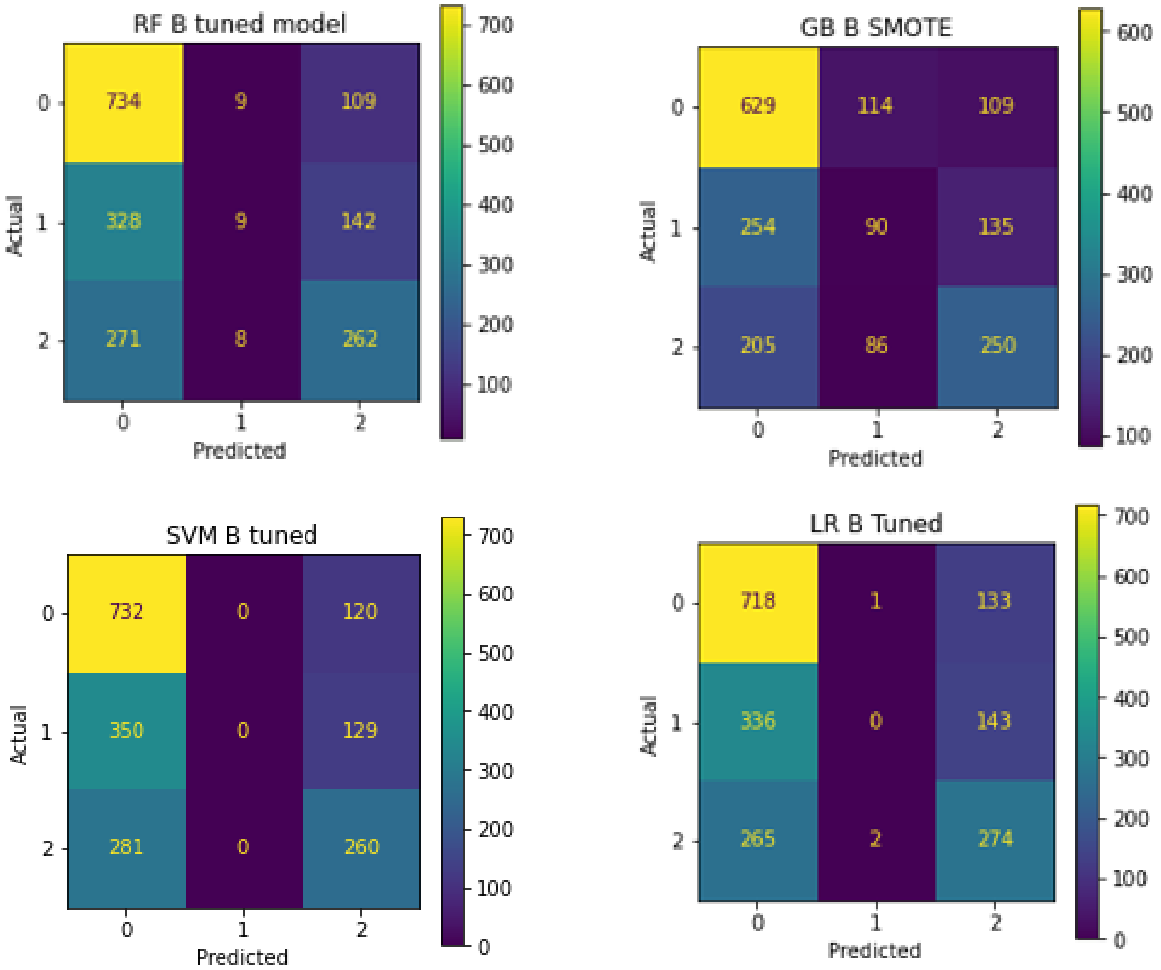


Figure 3. Confusion matrices of the most accurate Random Forest (upper left), Gradient Booster (upper right), Support Vector Machine (bottom left), and Logistic Regression (bottom right)

To further reflect on model performance, The best-performing model for each algorithm in terms of the weighted precision is displayed in table 5. Once again, a Random Forest model in combination with differential features turned out to be superior to other models, with a weighted precision of 52.6%. In contrast to the most accurate model, ClusterCentroids undersampling had been applied.

Table 5

Best performing models for each algorithm in terms of weighted precision

Algorithm	Feature set	Resampling	Weighted Precision
Random Forest	*Model B*	*Clustroids*	52.6%
Gradient Booster	*Model A*	*Clustroids*	50.6%
Support Vector Machine	*Model A*	*Balanced class weights*	50.5%
Logistic Regression	*Model B*	*Clustroids*	49.4%

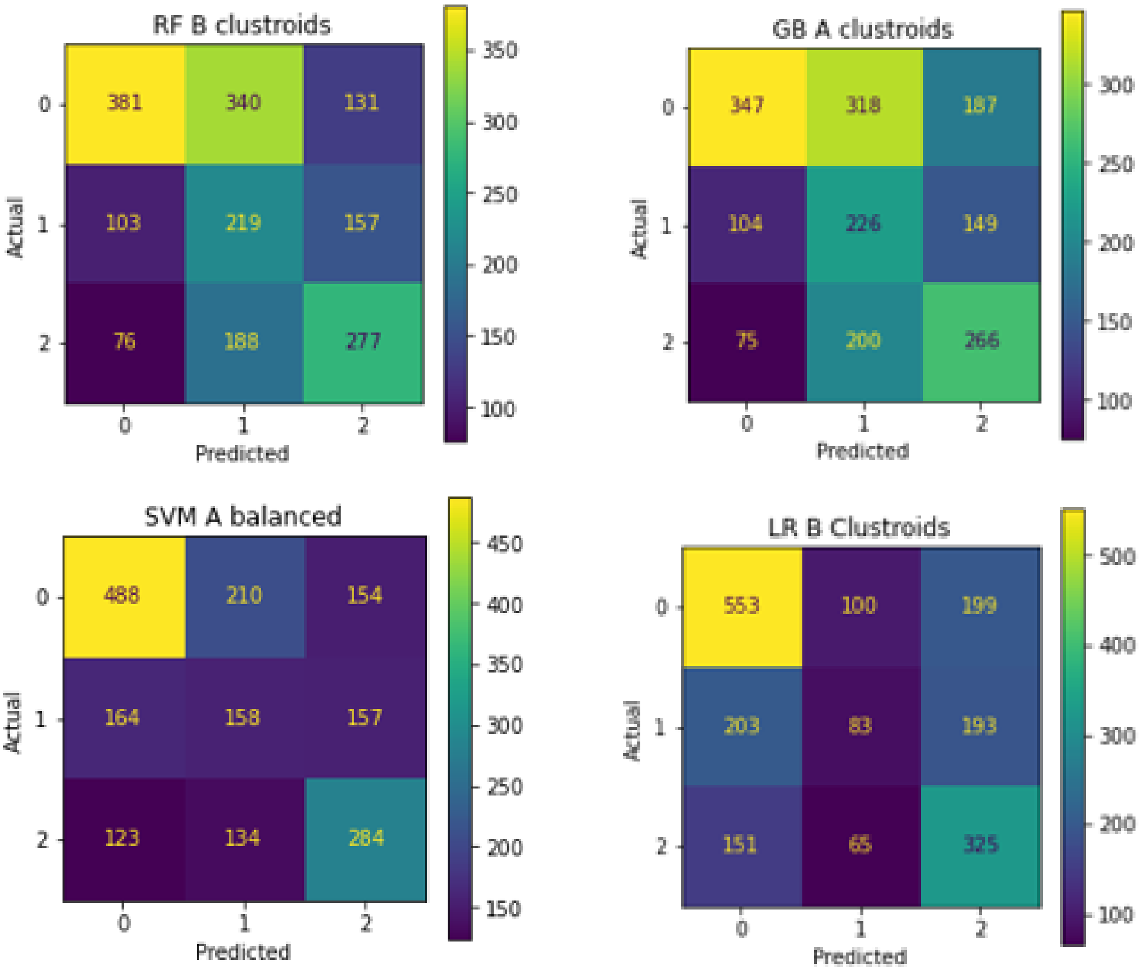


Figure 4. Confusion matrices of the most precise Random Forest (upper left), Gradient Booster (upper right), Support Vector Machine (bottom left), and Logistic Regression (bottom right)

Confusion matrices of the most precise models are visualized in figure 4 above. The application of ClusterCentroids undersampling seemed to be the best approach when aiming for a high detection rate of draws, although this seldom resulted in more precise draw classifications. Looking at the RF and GB models in combination with CC, a draw was the most predicted match outcome, which is noteworthy given the initial focus on this outcome without resampling.

Since many features in the final models showed high collinearity (see appendices 2.2, 2.3, and 2.4), an attempt was made to improve performance by reducing the number of variables. This reduction was done by Recursive Feature Elimination, with an RFECV model from the *Yellowbrick* library (see Figure 5). No noteworthy improvements were made after 25 variables. In the end, RFE did not lead to an increase in performance compared to the initial predictions. The 25 features with the highest importance in the most accurate- and precise model are displayed in table 6, and were selected based on their average decrease in impurity.

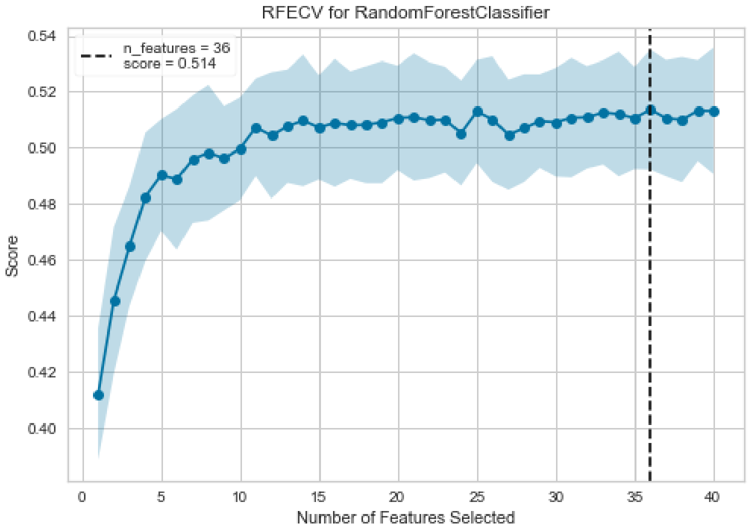


Figure 5. Recursive feature elimination of model B. A repeated stratified k-fold cross-validation was used. n_splits was set to 10, n_repeats to 3, and the performance to accuracy.

The optimal models showed great similarity in their predictor rankings. *FIFA ratings*, *ELO proxy*, *chances*, *goal difference*, *shots*, *possession*, *corners against*, and *percentage possible points* are among the top-10 features in both models.

Table 6

Feature importances from the most accurate (left) and precise (right) model

Features RF B *No resampling*	Feature importance	Features RF B *CC undersampling*	Feature importance
Overall rating	0.0774287181	Overall rating	0.0673909312
Defensive rating	0.0696782207	Optimized ELO proxy	0.0562757281
Optimized ELO proxy	0.0625132694	Attacking rating	0.0537612217
Attacking rating	0.0593460003	Defensive rating	0.053005637
Chances	0.0418236696	Chances	0.0421500781
Goal difference	0.0369458594	Average possession	0.0340534922
Average shots	0.0359852843	% Possible points	0.0339708492

Average possession	0.0341902205	Average shots	0.0332929165
Average corners against	0.0305026094	Goal difference	0.0328407005
% Possible points	0.0297447962	Average corners against	0.0317814305
Average corners	0.0285604145	Winning percentage	0.0294789326
Average shots off	0.0271110345	Average corners	0.0282531326
Winning percentage	0.0261149052	Aggression	0.0261881155
Average crosses	0.0253079624	Average shots off	0.0260889361
Aggression	0.0248557879	Precision	0.025278003
Finishing	0.0247354602	Average crosses	0.0252039596
Average age	0.0245912057	Finishing	0.0245567934
Precision	0.0242505037	Average BMI	0.0240219349
Average BMI	0.0240334149	Average height	0.0233938565
Average height	0.0235540998	Average age	0.0233703631
Clinical	0.0231190803	Clinical	0.0233384105
Average weight	0.0220522241	Average weight	0.0222313781
Losing percentage	0.0215460484	Current ranking	0.022177961
Current_ranking	0.0212944077	Losing percentage	0.0206867687
Players over 30	0.0191855419	Players over 30	0.0194821198

Key: Feature importance was determined by mean decrease in impurity

5.2 Selection of ELO Rating

The original- and proxy ELO calculations were tested in varying settings of K and home team advantage (HTA). The spearman's correlation between ELO scores and match outcomes was taken for each combination, as displayed in tables 7 and 8. Since both teams had separate ratings, the average was taken over their absolute correlation with the target variable. The optimal setting for the *original ELO* was K=25 and HTA = 0, whereas the highest correlation for the *ELO proxy* was achieved with K=10 and HTA = 200.

Table 7

Spearman's correlation between original ELO variants and match outcome

K	HTA = 0	HTA = 50	HTA = 100	HTA = 150	HTA = 200	Mean
10	0.22	0.22	0.219	0.218	0.217	0.2188
25	0.221	0.221	0.22	0.219	0.218	**0.2198**
50	0.22	0.22	0.219	0.218	0.218	0.2190
75	0.216	0.215	0.215	0.215	0.214	0.2150
100	0.212	0.21	0.21	0.21	0.211	0.2106
Mean	**0.2178**	0.2172	0.2166	0.2160	0.2156	0.2166

Table 8

Spearman's correlation between proxy ELO variants and match outcome

K	HTA = 0	HTA = 50	HTA = 100	HTA = 150	HTA = 200	Mean
10	0.243	0.243	0.243	0.243	0.243	**0.2430**
25	0.241	0.241	0.241	0.241	0.242	0.2412
50	0.234	0.233	0.234	0.234	0.235	0.2340
75	0.227	0.226	0.226	0.227	0.228	0.2268
100	0.22	0.22	0.22	0.22	0.221	0.2202
Mean	0.2330	0.2326	0.2328	0.2330	**0.2338**	0.2330

To observe the correlations in more detail, plots with correlations during particular stages of the season can be found in figures 6 and 7. The correlation of the *original ELO rating* is lower during the first 15 stages of the season compared to the *ELO proxy*, which is in line with existing literature. After this initial convergence period, all ratings showed a similar trend. Furthermore, the *optimized original ELO* ratings displayed a higher correlation with the target variable compared to the *original ELO* rating during the first 10 stages.

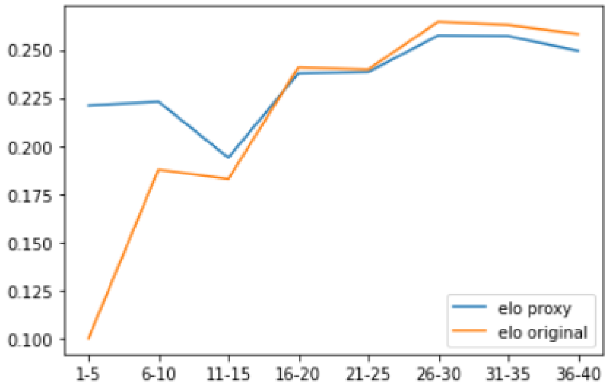


Figure 6. Comparison between ELO ratings witout optimizations and their correlation to match outcome

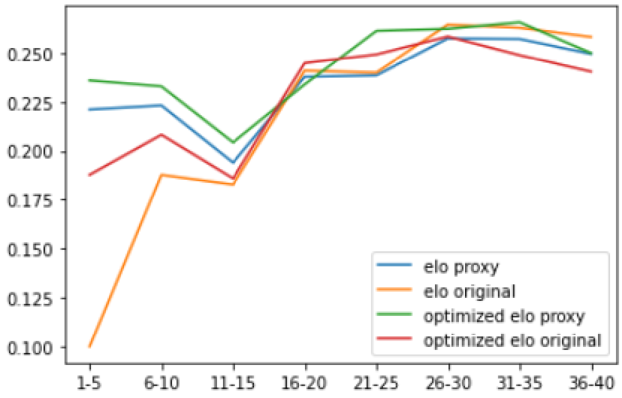


Figure 7. Comparison Between ELO ratings with optimizations and their correlation to match outcome

To finalize ELO comparisons, ratings were used in logistic regression models. The accuracy of each rating and comparison to baselines can be found in table 9. The *optimized ELO proxy* was the best performer with a predictive accuracy of 52.4% and was included in the final models. All ratings outperformed the majority baseline and LR model in Herbinet (2018). However, no model improved on bookmaker predictions solely on ELO rating.

Table 9

Performance of logistic regression models for different ELO operationalizations

Features	Accuracy	Majority baseline	Odds baseline	Herbinet (2018)
ELO original	49.5%	+4.0%	-3.6%	+0.9%
ELO proxy	51.1%	+5.7 %	-1.8%	+2.7%
Optimized ELO original	49.4%	+3.9%	-3.5%	+0.8%
Optimized ELO proxy	52.4%	+6.9 %	-0.6%	+3.9%

5.3 Selection of Match Event

A full overview of match events can be found in appendix 1.2. As mentioned, three different methods were compared: the average occurrence throughout the season, the absolute occurrence during the last game, and the average occurrence during the last five matches. RF models for each feature set were used to compare predictive abilities. In the end, events calculated as the weighted average occurrence over the entire season outperformed the other approaches, as can be seen in table 10.

Table 10

Performance of random forest models for different match event operationalizations

Match event type	Accuracy
Weighted averages season	50.6%
Past match occurrence	45.5%
Weighted average past 5 matches	48.5%

Since several events showed collinearity (see appendix 2.1), the number of events was reduced with RFE to find the optimal amount of events to retain. The outcome of the RFECV model from the *Yellowbrick* library has been visualized in figure 8. No noteworthy improvements were made after 10 features, hence only the 10 most informative events were included in the final models. Individual importance of events based on the mean decrease in impurity are visualized in figure 9. Match events such as *shots on and off target*, *possession*, *crosses* and *corners* were among the important events, whereas *fouls,* and *cards* contributed less.

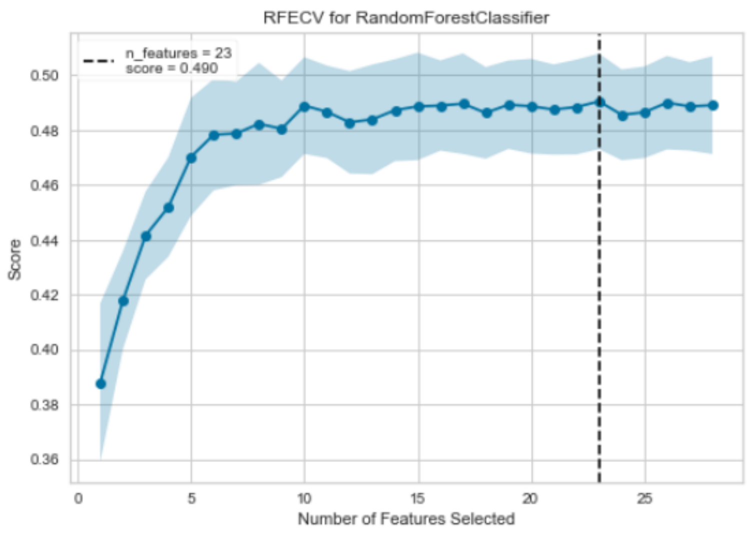


Figure 8. Recursive feature elimination of match events. A repeated stratified k-fold cross validation was used. n_splits was set to 10, N_repeats to 3, and the performance to accuracy.

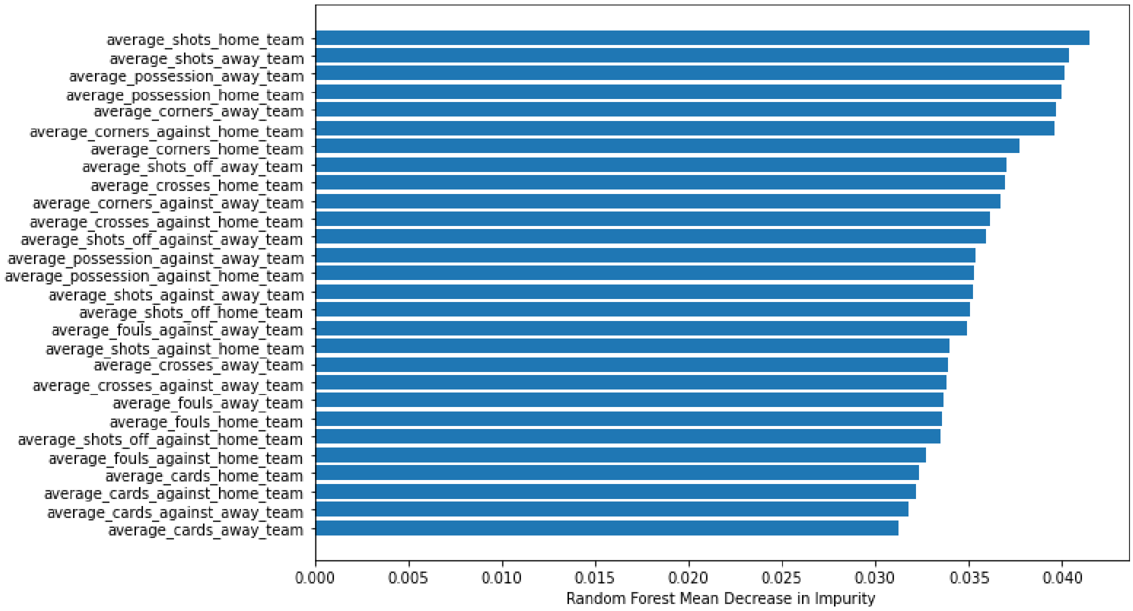


Figure 9. Mean decrease in impurity per match event in descending order

5.4 Full Variable Frames

All variables present in models A and B can be found in appendix 1.7 or the collinearity plots. Model B contains the same variables (apart from *formation* and *stage*) as Model A, but each variable is now represented by a differential score. In addition, the performance of all models on individual classes is displayed in appendices 1.4, 1.5 and 1.6

5.4.1 Random Forest

RF models trained on differential features outperformed their *model A* counterparts, regardless of resampling technique. Furthermore, the highest accuracy and precision during this thesis were both obtained using RF algorithms. Lastly, all RF models using some form of resampling showed increased weighted precision, but also a reduced overall accuracy. RF models combined with CC undersampling displayed excellent detection rates of ties up to 46%. Metrics of the RF models can be found in table 11.

Table 11

RF models classification metrics and performance on ties

Model	Accuracy model	Precision weighted	Precision draw	Recall draw	F1-score draw
RF A	52.9%	48.3%	0.35	0.01	0.02
RF A*	50.6%	50.4%	0.32	0.29	0.30
RF A**	49.4%	49.2%	0.30	0.27	0.28
RF A***	45.8%	51.0%	0.29	**0.43**	0.34
RF B	**53.7%**	48.7%	**0.35**	0.02	0.04
RF B*	51.3%	50.6%	0.32	0.27	0.29
RF B**	50.6%	50.9%	0.32	0.32	0.32
RF B***	46.8%	**52.6%**	0.29	0.46	**0.36**

Key: * = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling

5.4.2 Gradient Boosting

The performance metrics of GB models can be found in table 12. The GB was the only algorithm to put emphasis on ties without the use of resampling techniques, but also achieved the lowest overall accuracy compared to other classifiers. Furthermore, the accuracy of a GB somewhat increased when combined with SMOTE oversampling. As was the case with RF models, CC resampling led to excellent detection rates of draws up to 48% and resulted in the highest F1-scores obtained during this study.

Table 12

GB classification metrics and performance on ties

Model	Accuracy model	Precision weighted	Precision Draw	Recall draw	F1-score draw
GB A	51.1%	46.3%	0.29	0.09	0.13
GB A**	51.2%	47.0%	0.29	0.12	0.17
GB A***	44.8%	**50.1%**	0..30	0.47	**0.37**
GB B	51.1%	46.8%	0.27	0.12	0.17
GB B**	**51.8%**	48.9%	**0.31**	0.19	0.23
GB B***	41.9%	49.8%	0.28	**0.48**	0.35

Key: ** = SMOTE oversampling, *** = Clustroids undersampling

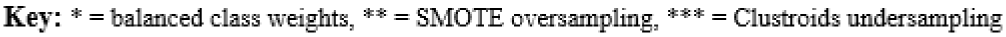
5.4.3 Support Vector Machine

The performance of SVM models can be found in table 13. As was the case with RF models, all models using differential features outperformed their counterparts. SVM algorithms entirely ignored a draw outcome without resampling. Looking at the F1 scores, this classifier seems to be most robust when it comes to resampling approach selection. The accuracy of the models once again did not improve by resampling, whereas the weighted precision did improve for all resampled models. The detection rate of draws by SVM models with CC undersampling was significantly lower compared to their RF and GB counterparts.

Table 13

SVM classification metrics and performance on ties

Model	Accuracy model	Precision weighted	Precision Draw	Recall draw	F1-score draw
SVM A	52.8%	39.0%	0.00	0.00	0.00
SVM A*	49.7%	**50.5%**	0.31	0.33	**0.32**
SVM A**	48.5%	49.3%	0.30	0.31	0.30
SVM A***	47.4%	49.3%	0.29	**0.35**	0.32
SVM B	**53.0%**	39.2%	0.00	0.00	0.00
SVM B*	49.8%	49.8%	**0.32**	0.30	0.31
SVM B**	49.1%	49.2%	0.32	0.31	0.31
SVM B***	48.9%	49.4%	0.31	0.32	0.31



Key: * = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling

5.4.4 Logistic Regression

The performance of the LR models can be found in table 14. The performance of models without resampling was almost identical to SVM models without resampling. From all resampled models, the LR trained on differential features (*model B*) and using class weights turned out to be the worst-performing model on ties, and simultaneously the best performer when it came to the overall accuracy. This trade-off between accuracy and performance on draws seems to be a general pattern among algorithms.

Table 14

Logistic Regression models classification metrics and performance on a draw

Model	Accuracy model	Precision weighted	Precision Draw	Recall draw	F1-score draw
LR A	53.0%	39.1%	0.00	0.00	0.00
LR A*	49.0%	49.0%	0.30	0.27	0.28

LR A**	47.3%	47.6%	0.27	0.25	0.26
LR A***	46.2%	47.9%	0.28	**0.32**	**0.29**
LR B	**53.0%**	39.2%	0.00	0.00	0.00
LR B*	52.6%	47.9%	0.31	0.06	0.10
LR B**	49.6%	47.4%	0.28	0.16	0.20
LR B***	51.3%	**49.4%**	**0.33**	0.17	0.23

Key: * = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling

6. Discussion

During this thesis, the aim was to explore to what extent the outcome of football matches could be predicted. In order to do so, feature operationalizations and algorithms were compared, the most important features were identified, and resampling techniques were utilized.

6.1 Match Outcome Predictions

In the end, only one approach outperformed the classifications made by bookmakers. An RF model with differential features (*model B*) and hyperparameters found in table 3 was able to add 0.7% accuracy, resulting in 53.7% correct classifications. This model however only detected 2% of all matches ending in a tie, while roughly 25% of games have this outcome. The inability to classify ties directly relates to the research question since it limits the extent to which most ML algorithms can be used for match outcome prediction. The most accurate model slightly underperformed compared to RF models in related studies by Tax and Joustra (2015) with 55.2% and Baboota and Kaur (2019) with 56.4%, while the performance metrics in Eryarsoy and Delen (2019) were deemed unrealistic and most likely resulting from noise, incorrect resampling, or missing data imputations. One crucial distinction between Tax and Joustra (2015) is their inclusion of bookmakers' data, whereas it was used as a baseline in this thesis. Furthermore, the minor gap in predictive capabilities could be caused by initial differences in the distribution of outcomes, with home team victories having a higher chance of being classified correctly. This was also the case when reflecting on the data used in Baboota and Kaur (2018), with a relatively small sample size (n=432) and over half the matches ending in home team victories. Two of their conclusions however were validated by the results of this study: differential features are a better fit compared to separate variables, and a GB algorithm is the best approach when interested in minority classification. Even though the GB models achieved the lowest accuracy, it showed great potential as it was the only algorithm to consider all outcomes without interventions to shift focus towards minority classes. This trade-off between accuracy and minority classification has been a recurrent theme in existing literature and was evident during this study. This makes it all the more surprising that the most accurate GB model was achieved in combination with SMOTE, while also predicting 15% of matches to end in a draw. This percentage closes in on the real-world distribution of outcomes and is encouraging for prospective improvements. Furthermore, GB in combination with CC undersampling made it possible to bias the classifier towards a draw, with recall scores close to 50%. As a next step, it is important to increase the precision of draw classifications, which were relatively unaffected by resampling. Reflecting back on the main research questions, it can be concluded that it is possible to predict football outcomes in a multi-class manner, but it does require the correct algorithm and type of intervention to reduce bias towards home team victories.

6.2 Feature Operationalizations and Importances

As for the usefulness of the ELO rating in association football, an approach was tried to overcome its convergence issues and reliance on data from previous seasons. Similar to Herbinet (2018), this thesis used the *European Soccer Database* to calculate a variation on the ELO rating as a metric for a football team's relative strength. The approach during this thesis outperformed their LR model by almost 4%. Looking at the feature importances in table 6, the *ELO proxy* ranked third and second in the most accurate- and precise model and was able to classify 52.4% of matches correctly without the inclusion of other features. The *ELO proxy* also outperformed the top-2 most important features in Eryarsoy and Delen (2019), which were *percentage possible points* and *current ranking*, strengthening the initial hypothesis that their accuracy of 76% was due to unrelated factors. On the other hand, the *ELO proxy* got outperformed by FIFA ratings as a metric for a team's relative strength. Which was deemed the most informative feature by both models, despite high collinearity with the *ELO proxy*. Both ratings also showed high correlations with obvious performance markers such as current ranking, winning- and losing percentages, percentage of possible points, goal difference, and form, as displayed in appendix 2.4. Since one of the goals was to reflect on the importance of features, it was decided not to exclude these variables but rather look at their individual performance for comparison. It is likely that similar performance can be achieved with only a fraction of these correlating features, given the model performance solely using the ELO scores that was only 1.3% off the most accurate model using a full variable frame. Lastly, the approach taken by Eryarsoy and Delen (2019) to quantify match events as weighted averages over all matches outperformed to other two proposed methods. Predictions solely on match events did not turn out to be most useful, but events like as *possession*, *shots*, *corners*, and *crosses* did rank among the top 15 features. The offensive occurrence of match events seemed to be more informative compared to the defensive occurrence when looking at table 6.

6.3 Limitations

Factors with proven effects on match outcome could not be taken into consideration, such as games outside the national competition (Verheijen et al., 2012), a team's manager (Besters et al., 2016), and players outside the starting line-up (Varela-Quintana et al., 2016). Furthermore, the dataset was relatively small and dates back 6 years. In those years, the game of football has been subject to various changes (Connelly, 2020), which could make some of the conclusions already outdated. The biggest limitation of this study is the way algorithms made their predictions. Based on the individual feature importances in table 6, the algorithms mostly relied on obvious performance markers, which brings about predictions that always favor the stronger team, rather than finding scenarios where the weaker team has an increased chance of victory.

6.4 Feature Research

To address the final limitation, a follow-up on this study could be the exclusive use of matches in which the weaker team ended up victorious, and apply a variety of clustering algorithms to this data. This might reveal the existence of patterns within those matches, which could improve the flexibility of decision-making making beyond picking the strongest team. Furthermore, the performance of FIFA ratings during this study makes a great case to further investigate these ratings. During this study, only the overall rating of players was used. Ratings also contain more detailed characteristics such as physical, athletic, mental, and skill aspects to express a player’s qualities. A more in-depth use of the ratings could prove worthwhile. Lastly, to reflect on the algorithm’s ability to learn throughout a season, a distinction could be made between matches played at various stages.

7. Conclusion

Due to the lack of studies predicting football matches, there was still no clear consensus as to what methodology and operationalizations to use, which features to adopt, and how to address the poor performance on ties. By applying a comparative approach, several of these unknowns have been addressed. Ensemble methods such as the RF and GB were the most useful classifiers in varying scenarios. Differential features mostly outperformed separate features for home- and away teams while simultaneously reducing dimensions and overlap. Resampling approaches in combination with ensemble methods can significantly improve the detection of lesser occurring match outcomes but had little effect on the precision of classifications. Finally, a full overview of features and their importance was given to further reflect on how decisions came about. Findings during this study can serve as a benchmark for future research, and help select the appropriate methodology when there is a preference for certain performance measures. The majority of previous studies suffered from under-reporting, which made it hard to replicate and reflect upon their findings. Moreover, this study is one of the few studies in this field to make use of a dataset collected over more than one league, and the first to compare various methods to improve minority classification. In contrast to scientific relevance, the predictions during this study are currently not accurate enough and too difficult to obtain to generate direct societal impact.

7. References

Aalbers, B., & Van Haaren, J. (2018, September). Distinguishing between roles of football players in play-by-play match event data. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 31-41). Springer, Cham

Aldous, D. (2017). Elo ratings and the sports model: A neglected topic in applied probability? *Statistical Science*, *32*(4), 616-629.

Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, *21*, 449-470.

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741-755.

Besters, L. M., van Ours, J. C., & van Tuijl, M. A. (2016). Effectiveness of in-season manager changes in English Premier League Football. *De Economist*, *164*(3), 335-356.

Bex, T (2021). Powerful Feature Selection with Recursive Feature Elimination (RFE) of Sklearn. *Towards Data Science.*

Bisberg, A. J., & Cardona-Rivera, R. E. (2019, October). Scope: Selective cross-validation over parameters for elo. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 15, No. 1, pp. 116-122).

Buchdahl, J. (2003). *Fixed odds sports betting: Statistical forecasting and risk management*. Summersdale Publishers LTD-ROW.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, H. (2019). Neural network algorithm in predicting football match outcome based on player ability index. *Advances in Physical Education*, *9*, 215-222.

Connelly, B. (2020, April 19). How soccer has changed in the past 10 years: From Mourinho’s peak to reign of superclubs. *ESPN*. https://www.espn.com/soccer/english-premier-league/story/4086497/how-soccer-has-changed-in-the-past-10-years-from-mourinhos-peak-to-reign-of-super-clubs

Constantinou, A., & Fenton, N. (2017). Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, *124*, 93-104.

Dabbura, I. (2018). K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. *Towards Data Science*.

Dahinden, C., & Ethz (2011). An improved Random Forests approach with application to the performance prediction challenge datasets. *Hands-on Pattern Recognition, Challenges in Machine Learning*, *1*, 223-230.

Elo, A. E. (1978). *The rating of chessplayers, past and present.* BT Batsford Limited

Eryarsoy, E., & Delen, D. (2019). Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods.

Farid, D. M., Nowé, A., & Manderick, B. (2016). A new data balancing method for classifying multiclass imbalanced genomic data. In 25th Belgian-Dutch Conference on Machine Learning (Benelearn) (pp. 1-2).

FIFA (2021, August 30). Report on ten years of international transfers during the 2011-2020 period. https://www.fifa.com/legal/media-releases/fifa-publishes-report-on-ten-years-of-international-transfers

Forslund, M. (2017). Innovation in soccer clubs–the case of Sweden. Soccer & Society, 18, 374-395.

Garnica-Caparrós, M., & Memmert, D. (2021). Understanding gender differences in professional European football through machine learning interpretability and match actions data. *Scientific Reports*, *11*(1), 1-14.

Gandhi, R. (2018). Support Vector Machine – Introduction to Machine Learning Algorithms. *Towards Data Science.*

Goddard, J. (2006). Who wins the football?. *Significance*, *3*(1), 16-19.

Graham, I., & Scott, H. (2008). Predicting bookmakers odds and efficiency for UK football. *Applied economic*, 40(1), 99-109.

Harper, J (2021, March 5). Data experts are becoming football's best signings. *BBC*. https://www.bbc.com/news/business-56164159

Herbinet, C. (2018). Predicting football results using machine learning techniques. *MEng thesis, Imperial College London.*

Heuer, A., & Rubner, O. (2009). Fitness, chance, and myths: an objective view on soccer results. *The European Physical Journal B*, *67*(3), 445-458.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.

Horning, N. (2010, December). Random Forests: An algorithm for image classification and generation of continuous fields data sets. *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan*, 910, 1-6.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of forecasting*, *26*(3), 460-470.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54.

Kaggle European Soccer Database [https://www.kaggle.com/hugomathien/ soccer12].

Kaggle European Soccer Database Supplementary [https://www.kaggle.com/datasets/jiezi2004/soccer].

Keogh, F., & Rose, G. (2013, October 3). Football betting – the global gambling industry worth billions. *BBC*. https://www.bbc.com/sport/football/24354124

Kho, J. (2018). Why Random Forest is my favorite machine learning model. *Towards Data Science.*

Kish, L. (2017). Some statistical problems in research design. In *Research Design* (pp. 64-78). Routledge.

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. *International Journal of Forecasting*, *36*(4), 1329-1341.

Krifa, A., Spinelli, F., & Junca, S. (2021). *On the convergence of the Elo rating system for a Bernoulli model and round-robin tournaments* (Doctoral dissertation, Université Côte D'Azur).

Last, F., Douzas, G., & Bacao, F. (2017). Oversampling for imbalanced learning based on k-means and smote. arXiv preprint arXiv:1711.00837.

Macdonald, B. (2012, March). An expected goals model for evaluating NHL teams and players. In *Proceedings of the 2012 MIT Sloan Sports Analytics Conference*.

Matano, F., Richardson, L. F., Pospisil, T., Eubanks, C., & Qin, J. (2018). Augmenting adjusted plus-minus in soccer with FIFA ratings. *arXiv preprint arXiv:1810.08032*.

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20-33.

Muralidhar, K. (2021). The right way of using SMOTE with Cross-validation. *Towards Data Science*.

Obadia, Y (2017). The use of KNN for missing values. *Towards Data Science*

Prasetio, D. (2016, August). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE.

Raj, J. (2020, November 7). Perfect Recipe for Classification Using Logistic Regression. *Towards Data Science.*

Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017, August). Predicting football matches results using Bayesian networks for English Premier League (EPL). In *Iop conference series: Materials science and engineering* (Vol. 226, No. 1, p. 012099). IOP Publishing.

Reza, M. S., & Ma, J. (2018, August). Imbalanced histopathological breast cancer image classification with convolutional neural network. In *2018 14th IEEE International Conference on Signal Processing (ICSP)* (pp. 619-624). IEEE.

Riezebos, M., van de Velden, M., Birbil, S. I., BV, S. O. S., van der Knaap, R., & Talsma, B. (2021). Identifying Play Styles of Football Players Based on Match Event Data.

Rohde, M., & Breuer, C. (2016). The financial impact of (foreign) private investors on team investments and profits in professional football: Empirical evidence from the premier league. Applied Economics and Finance, 3, 243-255.

Swaminathan, S. (2018). Logistic Regression – Detailed Overview. *Towards Data Science*

Seo, J. H., & Kim, Y. H. (2018). Machine-learning approach to optimize smote ratio in class imbalance dataset for intrusion detection. *Computational intelligence and neuroscience*, *2018.*

Tarbani, N. (2021). How the Gradient Boosting Algorithm works? *Analytics Vidhya*

Tax, N., & Joustra, Y. (2015). Predicting the Dutch football competition using public data: A machine learning approach. *Transactions on knowledge and data engineering*, *10*(10), 1-13.

Teli, L. K., Zaveri, N., & Shinde, P. (2018). Prediction of football match score and decision making process. *International Journal on Recent and Innovation Trends in Computing and Communication*, *6*(2), 162-165.

Tharwat, A. (2016). Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition*, *3*(3), 197-240.

Varela-Quintana, C., del Corral Cuervo, J., & Prieto-Rodriguez, J. (2016). The effect of an additional substitution in association football. Evidence from the Italian Serie A. *Revista de psicología del deporte*, *25*(1), 101-105.

Verheijen, R. (2012). Study on recovery days. *World Football Academy.*

World Football Elo Ratings. Elo ratings. http://eloratings.net/about

Yagci, H. (2021). Under-Sampling Methods for Imbalanced Data (ClusterCentroids, RandomUnder Sampler, NearMiss). *Towards Data Science.*

Yezus, A. (2014). Predicting outcome of soccer matches using machine learning. *Saint-Petersburg University*.

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., & Goyal, H. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, *6*(11).

Zhang, Y. P., Zhang, L. N., & Wang, Y. C. (2010, September). Cluster-based majority under-sampling approaches for class imbalance learning. In *2010 2nd IEEE International Conference on Information and Financial Engineering* (pp. 400-404). IEEE.

Appendix

1. Tables

Table 1.1

Matches per league and season

League	Country	Range	n
Premier League	England	2008-2016	3040
Bundesliga	Germany	2014-2016	612
Primera Division	Spain	2014-2016	760
Serie A	Italy	2014-2016	759
Ligue 1	France	2014-2016	760
Eredivisie	Netherlands	2015-2016	306

Table 1.2

Description of match events

Match event	Description
Card	Occurs when a rule is violated (foul) to a degree that the referee decides to further discipline the offender. Once a player receives two cards he has to leave the game without the option of being substituted.
Corner	Occurs when a ball crosses the backline while the last touch of the ball came off a player defending the goal on that side of the field. The attacking team can resume the play in the corners connected to the backline that was crossed.
Cross	Occurs when a player passes the ball towards the opposing team's goal with the intention to create a scoring opportunity for one of his teammates.
Foul	Occurs when the referee decides the actions of one- or more players are in violation of the rules. Can be supplemented by a card for the offence, but this is not necessary to constitute a foul
Possession	Displays the percent of the time in which a team had possession over the ball during a match compared to total match-time.
Shot on target	Shot that would have resulted in a goal if untouched
Shots off target	Shot that would not have resulted in a goal if untouched

Key: match events were operationalized both as offensive (achieved) and defensive (conceded) occurrence of events.

Table 1.3

Description of hyperparameters functions

Algorithm	Hyperparameter	Description
Random Forest	*Max_depth*	Value to limit/increase the maximum depth of a single tree
	Max_features	Value to limit/increase the maximum number of features to consider when looking for the best data split
	Min_samples_split	The minimum amount of samples allowed in a leaf node
	N_estimators	Amount of singular trees to build before class selection by majority vote
Gradient Booster	*Learning_rate*	Controls the influence of individual trees
	Max_depth	Value to limit/increase the maximum depth of a single tree
	N_estimators	Amount of singular trees to build in a serial manner
	Subsample	The ratio of samples to fit individual base models
Support Vector Machine	*C*	Regularization parameter that adds a penalty for misclassifications
	Gamma	Determines how similar data points have to be in terms of location to be classified in the same class
	Kernel	Can be used to transform input from lower to higher dimensional spaces
Logistic Regression	*C*	Regularization parameter that adds a penalty for misclassifications
	Solver	different algorithms that can be used to optimize the weights of all variables in the model.

Table 1.4

Predictions per model on home team wins

Model	Precision	Recall	F1-score	Accuracy model
RF A	0.54	0.87	0.66	0.53
RF A*	0.63	0.62	0.62	0.51
RF A**	0.61	0.60	0.61	0.49
RF A***	0.66	0.46	0.54	0.46
RF B	0.55	0.86	0.67	0.54
RF B*	0.62	0.63	0.63	0.51
RF B**	0.63	0.60	0.61	0.51
RF B***	0.68	0.45	0.54	0.47
GB A	0.55	0.80	0.65	0.51
GB A**	0.57	0.76	0.65	0.51
GB A***	0.66	0.41	0.50	0.45
GB B	0.56	0.77	0.65	0.51
GB B**	0.58	0.74	0.65	0.52
GB B***	0.67	0.35	0.46	0.42
SVM A	0.53	0.87	0.66	0.53
SVM A*	0.63	0.57	0.60	0.50
SVM A**	0.62	0.56	0.59	0.48
SVM A***	0.62	0.53	0.57	0.47
SVM B	0.54	0.86	0.66	0.53
SVM B*	0.62	0.60	0.61	0.50
SVM B**	0.60	0.59	0.60	0.49
SVM B***	0.61	0.57	0.59	0.49
LR A	0.55	0.81	0.66	0.52
LR A*	0.62	0.59	0.60	0.49
LR A**	0.61	0.57	0.59	0.47
LR A***	0.61	0.52	0.56	0.46
LR B	0.54	0.84	0.66	0.53
LR B*	0.58	0.75	0.66	0.52
LR B**	0.60	0.63	0.61	0.50
LR B***	0.61	0.65	0.63	0.51

Key: ** = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling*

Table 1.5

Predictions per model on draws

Model	Precision	Recall	F1-score	Accuracy model
RF A	0.35	0.01	0.02	0.53
RF A*	0.32	0.29	0.30	0.51
RF A**	0.30	0.27	0.28	0.49
RF A***	0.29	0.43	0.34	0.46
RF B	0.35	0.02	0.04	0.54
RF B*	0.32	0.27	0.29	0.51
RF B**	0.32	0.32	0.32	0.51
RF B***	0.29	0.46	0.36	0.47
GB A	0.29	0.09	0.13	0.51
GB A**	0.29	0.12	0.17	0.51
GB A***	0.30	0.47	0.37	0.45
GB B	0.27	0.12	0.17	0.51
GB B**	0.31	0.19	0.23	0.52
GB B***	0.28	0.48	0.35	0.42
SVM A	0.00	0.00	0.00	0.53
SVM A*	0.31	0.33	0.32	0.50
SVM A**	0.30	0.31	0.30	0.48
SVM A***	0.29	0.35	0.32	0.47
SVM B	0.00	0.00	0.00	0.53
SVM B*	0.32	0.30	0.31	0.50
SVM B**	0.32	0.31	0.31	0.49
SVM B***	0.31	0.32	0.31	0.49
LR A	0.23	0.05	0.08	0.52
LR A*	0.30	0.27	0.28	0.49
LR A**	0.27	0.25	0.26	0.47
LR A***	0.28	0.32	0.29	0.46
LR B	0.00	0.00	0.00	0.53
LR B*	0.31	0.06	0.10	0.52
LR B**	0.28	0.16	0.20	0.50
LR B***	0.33	0.17	0.23	0.51

Key: ** = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling*

Table 1.6

Prediction metrics per model on away team wins

Model	Precision	Recall	F1-score	Accuracy model
RF A	0.51	0.45	0.48	0.53
RF A*	0.47	0.52	0.50	0.51
RF A**	0.47	0.52	0.50	0.49
RF A***	0.47	0.48	0.47	0.46
RF B	0.51	0.48	0.50	0.54
RF B*	0.48	0.55	0.51	0.51
RF B**	0.50	0.52	0.51	0.51
RF B***	0.49	0.51	0.50	0.47
GB A	0.48	0.43	0.46	0.51
GB A**	0.47	0.47	0.47	0.51
GB A***	0.44	0.49	0.47	0.45
GB B	0.49	0.45	0.47	0.51
GB B**	0.51	0.46	0.48	0.52
GB B***	0.43	0.48	0.45	0.42
SVM A	0.51	0.45	0.48	0.53
SVM A*	0.48	0.52	0.50	0.50
SVM A**	0.46	0.52	0.49	0.48
SVM A***	0.46	0.50	0.48	0.47
SVM B	0.51	0.48	0.50	0.53
SVM B*	0.47	0.52	0.49	0.50
SVM B**	0.47	0.50	0.48	0.49
SVM B***	0.47	0.51	0.49	0.49
LR A	0.50	0.50	0.50	0.52
LR A*	0.46	0.53	0.49	0.49
LR A**	0.45	0.51	0.48	0.47
LR A***	0.45	0.50	0.48	0.46
LR B	0.50	0.51	0.50	0.53
LR B*	0.46	0.58	0.51	0.52
LR B**	0.45	0.58	0.51	0.50
LR B***	0.45	0.60	0.52	0.51

Key: ** = balanced class weights, ** = SMOTE oversampling, *** = Clustroids undersampling*

Table 1.7

Description of variables, their calculation and related papers

Variable	Description	Source
Aggression	Calculated as the accumulated amount of cards within a season divided by the number of total fouls	Baboota & Kaur (2019)
Amount of formations	The number of formations tried throughout the season	Eryarsoy & Delen (2019)
Attack rating	Average of the overall FIFA ratings based on attackers and midfielders positioned on the flanks	Baboota & Kaur (2019);
Average age	The average age of all players in starting line-up	
Average BMI	The average BMI of all players in starting line-up (kg/m^2)	
Average height	The average height of all players in starting line-up	
Average weight	The average weight of all players in starting line-up	
Chances	Calculated over each match, by dividing the total shots from a team by the total amount of shots during the match. Reflects which team had the most scoring opportunities	MacDonald, 2012
Changes starting 11	Changes in a team's starting line-up compared to the previous match.	
Clinical	Shooting metric calculated by dividing the total amount of goals scored divided by the total amount of shots on target	MacDonald, 2012
Current ranking	Position in the league table. Displays all teams in descending order based on points, accumulated goal difference, and accumulated goals scored, in that sorting order	Eryarsoy & Delen (2016)
Draw percentage	Amount of games tied by a team divided by the total amount of games played by a team up until that point of the season	Tax & Joustra (2015)
Draw streak	Uninterrupted streak of ties	Tax & Joustra (2015)
Defensive rating	Average of the overall FIFA rating based on defenders and central midfielders	Baboota & Kaur (2019)

ELO rating	Reflection of a team's relative strength compared to other competitors in the same competition. The exact calculations can be found in paragraph 4.5	Herbinet (2018)
Finishing	Shooting metric reflecting the average amount of goals scored for each shot taken, includes both shots on- and off-target.	
Form	Calculated over 5 matches (win=2, draw=1, loss=0). Sum is divided by 10. For the first 4 games, sum was divided by matches played multiplied by 2.	Yezus (2014); Baboota & Kaur (2019)
Form away games	Calculated similarly to regular form, but only over the last 5 away games. Not updated during home games	
Form home games	Calculated similarly to regular form, but only over the last 5 home games. Not updated during away games	
Formation	Amount of keepers, defenders, midfielders, and attackers within a team's starting line-up. leads to values like 1442, 1443, and 1451, which was later one-hot encoded.	Eryarsoy & Delen (2016)
Goal difference	Total amount of goals scored in a season so far subtracted by the total amount of goals against. See Yezus (2014) for the exact scaling formula	Yezus (2014)
Losing percentage	Amount of games lost by a team divided by the total amount of games played up until that point	Tax & Joustra (2015)
Losing streak	Uninterrupted streak of games lost	Tax & Joustra (2015)
Overall rating	Average of the overall FIFA rating of players in the starting line-up	Baboota & Kaur (2019)
Percentage possible points	Percentage of how many points a team has thus far in a season relative to the maximum point possible	Eryarsoy & Delen (2016)
Players over 30 age	Amount of players over 30 years old in starting line-up	Eryarsoy & Delen (2016)
Players over 190 cm	Amount of players taller than 190 cm in starting lineup	

Precision	Shooting metric also known as Fenwick rating. Calculated by the accumulated shots on target divided by the total amount of shots	MacDonald, 2012
Promotion status	Displays if a team is promoted from a lower division. Can either be a 0 (not promoted) or 1 (promoted).	
Ratio of 4 most used players	Ratio that displays how many of the four most-used players so far are present in the team line-up. Sum is taken for all four players (present = 1, absent = 0) and divided by 4	
Score difference	Difference in goals scored within a match divided by the maximum difference during that stage of the season. See Yezus (2014) for the exact scaling formula	Yezus (2014); Baboota & Kaur (2019)
Stage	Round of the competition in which the game took place	
Top creator plays	Is either 0 or 1 and displays if the player with the most goal assists for his team thus far in a season is within the starting line-up	Tax & Joustra (2015)
Topscorer plays	Value of either 0 or 1 that displays if the player with the most goals for his team thus far in a season is within the starting line-up	Tax & Joustra (2015)
Winning percentage	Amount of games won by a team divided by the total amount of games played	Tax & Joustra (2015)
Winning streak	Uninterrupted streak of games won	Tax & Joustra (2015)

2. Figures

Figure 2.1

Correlations between match events

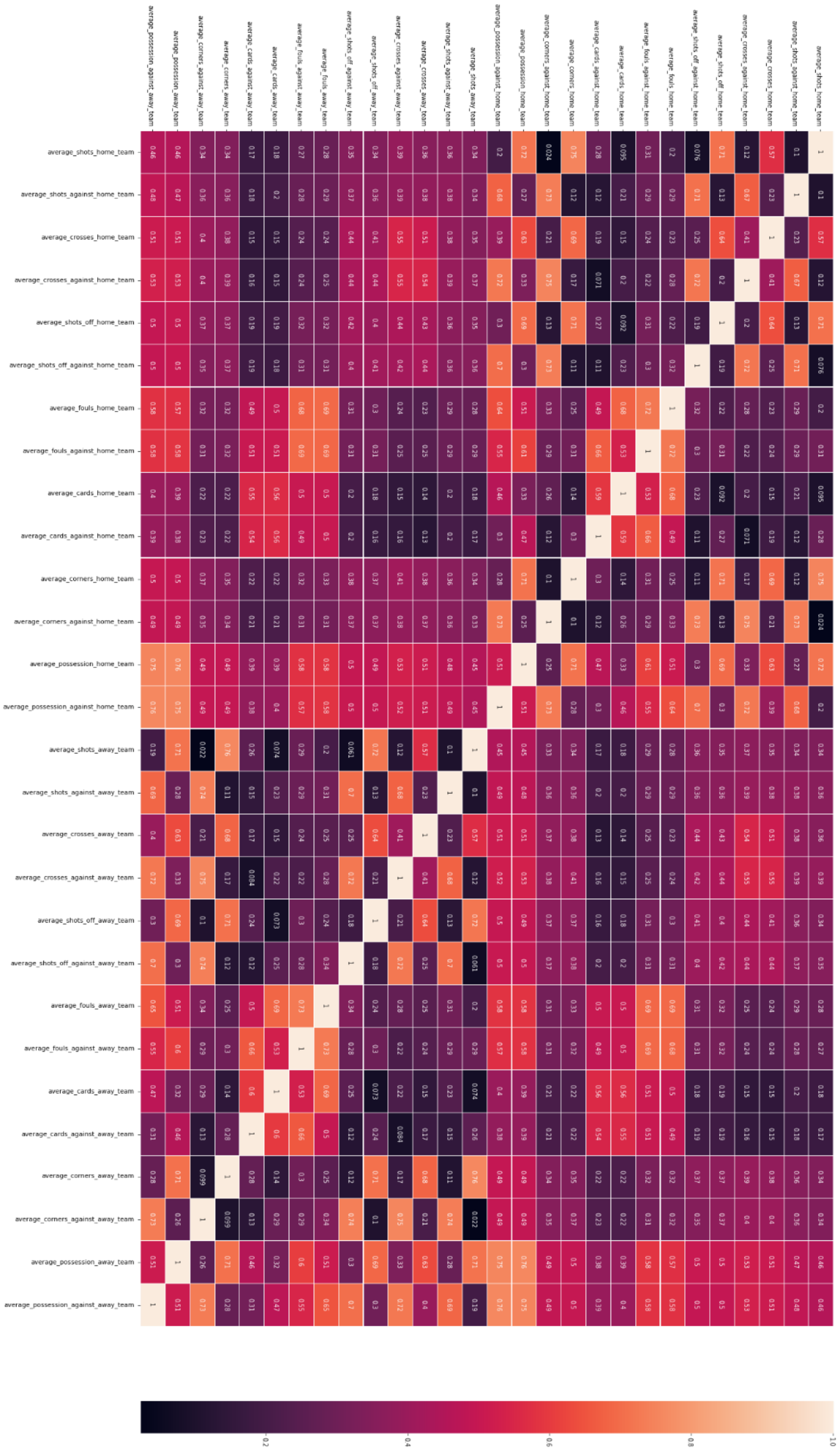


Figure 2.2

Correlations between home team variables

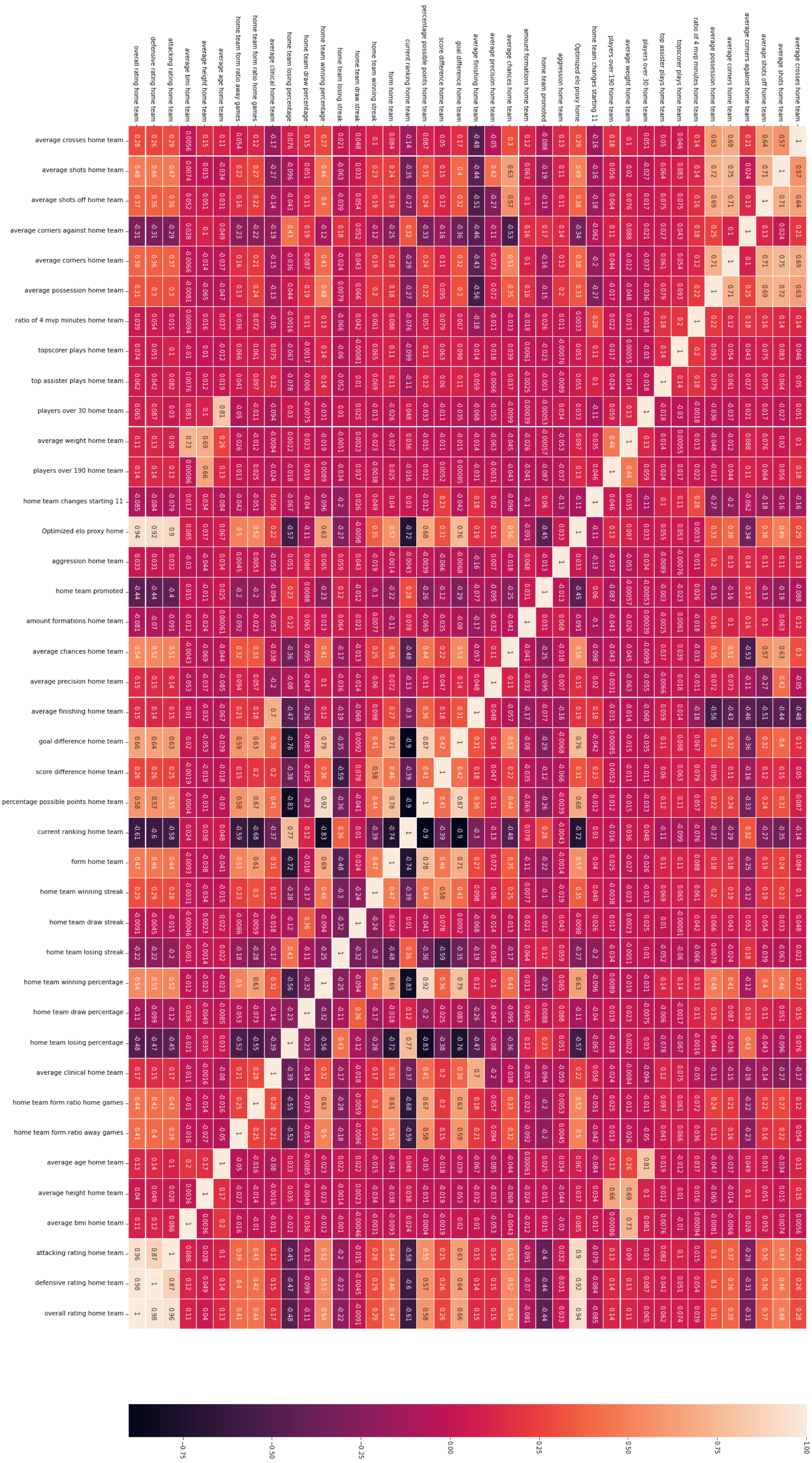


Figure 2.3

Correlations between away team variables

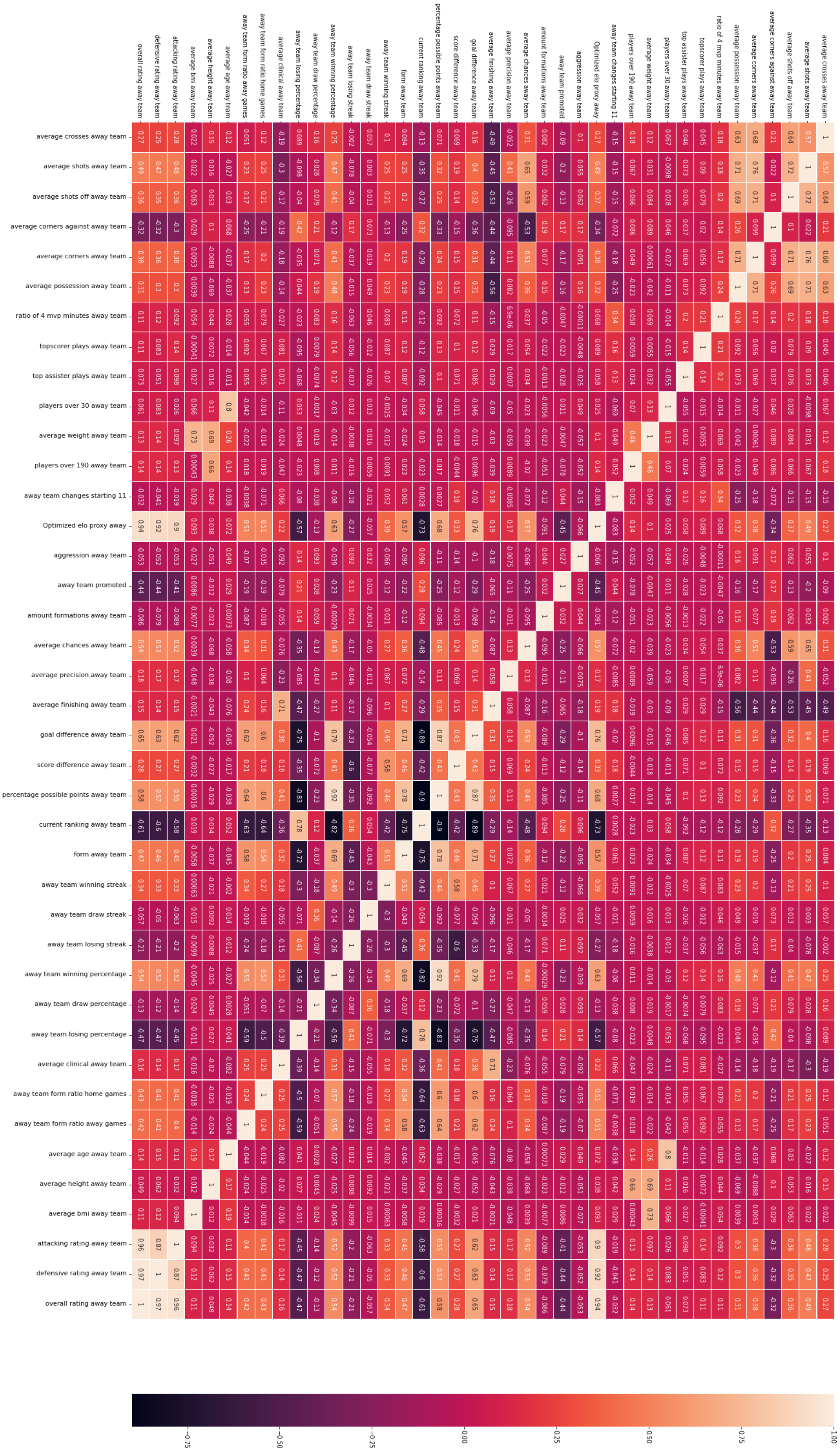
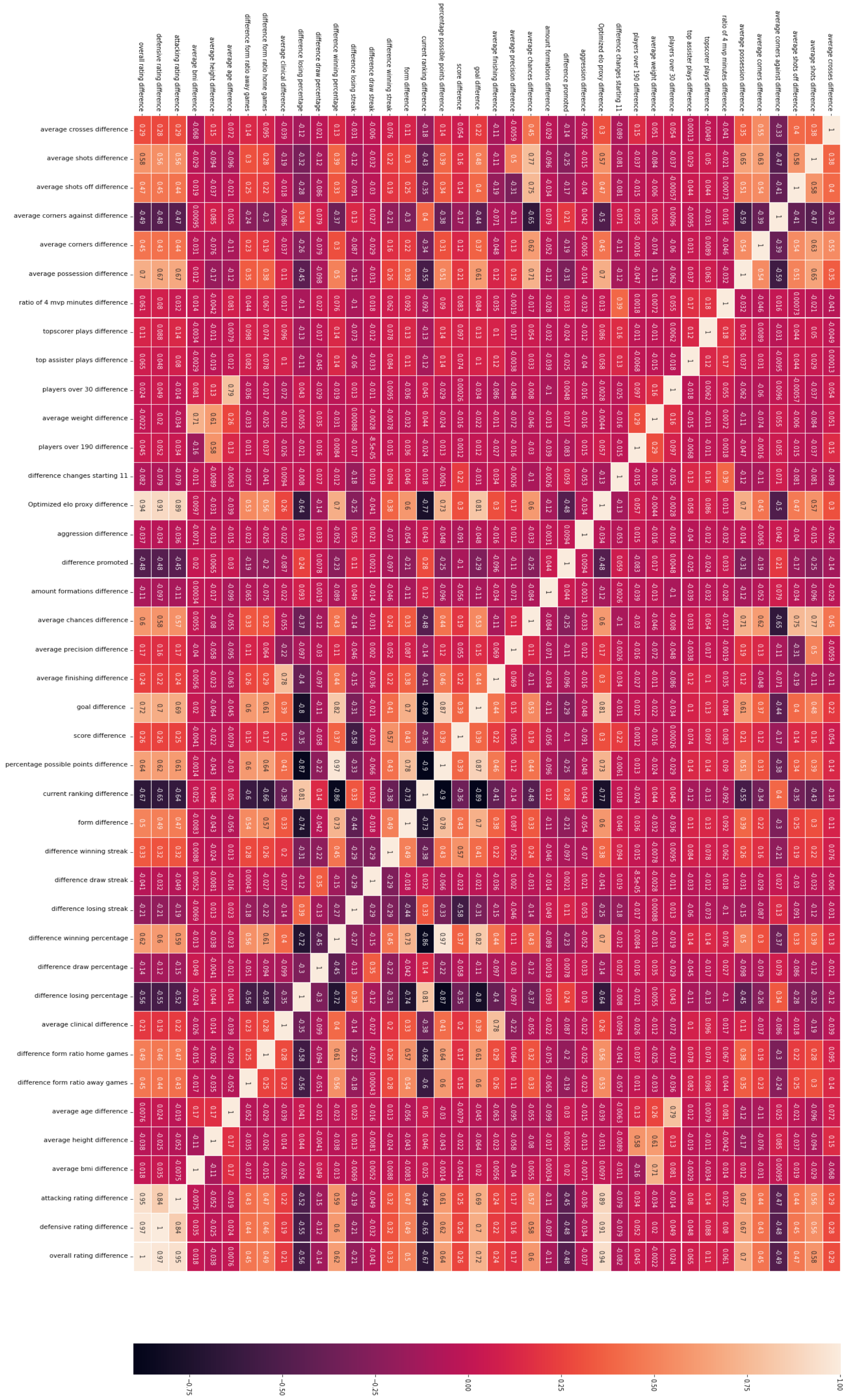


Figure 2.4

Correlations between differential variables


average crosses difference
average shots difference
average shots off difference
average corners against difference
average corners difference
average possession difference
ratio of 4 mvp minutes difference
topscorer plays difference
top assister plays difference
players over 30 difference
average weight difference
players over 190 difference
difference changes starting 11
Optimized elo proxy difference
aggression difference
difference promoted
amount formations difference
average chances difference
average precision difference
average finishing difference
goal difference
score difference
percentage possible points difference
current ranking difference
form difference
difference winning streak
difference draw streak
difference losing streak
difference winning percentage
difference draw percentage
difference losing percentage
average clinical difference
difference form ratio home games
difference form ratio away games
average age difference
average height difference
average bmi difference
attacking rating difference
defensive rating difference
overall rating difference