applied sciences

MDPI

Article

Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer

Youri Geurkink [1], Jan Boone [1,2], Steven Verstockt [3,†] and Jan G. Bourgois [1,2,*,†]

[1] Department of Movement and Sports Sciences, Ghent University, 9000 Ghent, Belgium; youri.geurkink@ugent.be (Y.G.); jan.boone@ugent.be (J.B.)
[2] Center of Sports Medicine, Ghent University Hospital, 9000 Ghent, Belgium
[3] Department of Electronics and Information Systems, Research Group IDLab, Ghent University—IMEC, 9052 Ghent, Belgium; steven.verstockt@ugent.be
* Correspondence: jan.bourgois@ugent.be
† These authors share last authorship.

Abstract: This study aimed to identify the strongest predictive variables of winning and losing in the highest Belgian soccer division. A predictive machine learning model based on a broad range of variables (n = 100) was constructed, using a dataset consisting of 576 games. To avoid multicollinearity and reduce dimensionality, Variance Inflation Factor (threshold of 5) and BorutaShap were respectively applied. A total of 13 variables remained and were used to predict winning or losing using Extreme Gradient Boosting. TreeExplainer was applied to determine feature importance on a global and local level. The model showed an accuracy of 89.6% ± 3.1% (precision: 88.9%; recall: 90.1%, f1-score: 89.5%), correctly classifying 516 out of 576 games. Shots on target from the attacking penalty box showed to be the best predictor. Several physical indicators are amongst the best predictors, as well as contextual variables such as ELO -ratings, added transfers value of the benched players and match location. The results show the added value of the inclusion of a broad spectrum of variables when predicting and evaluating game outcomes. Similar modelling approaches can be used by clubs to identify the strongest predictive variables for their leagues, and evaluate and improve their current quantitative analyses.

Keywords: association football; performance; performance analysis; KPI; game result

check for updates

Citation: Geurkink, Y.; Boone, J.; Verstockt, S.; Bourgois, J.G. Machine Learning-Based Identification of the Strongest Predictive Variables of Winning and Losing in Belgian Professional Soccer. *Appl. Sci.* **2021**, *11*, 2378. https://doi.org/10.3390/app11052378

Academic Editor: Rocco Furferi

Received: 9 February 2021
Accepted: 3 March 2021
Published: 8 March 2021

1. Introduction

The numerous unique chains of dynamic interactions between players during soccer games can be reduced to different performance indicators, to allow more insight into game performance [1]. A performance indicator is a selection, or combination, of action variables aiming to define some or all aspects of performance, and can be used to assess the performances of an individual, a team or the elements of a team [2]. Generally, sports performance judgments are prone to bias [3], as good outcomes are attributed to internal causes and bad outcomes to external causes [4]. Better insights into performance indicators could therefore not only aid the decision-making process of coaches and players with regard to training and game preparation, but also help other stakeholders related to the team, such as scouting and management, to correctly evaluate team and player performances [5].

Previously, several performance indicators have already been linked to performance in soccer, such as ball possession [6–9], number of passes [10,11], number of shots [6,8,10,12], number of shots on target [6,8,12], entries into the penalty box [9,13] and successfulness in duels [9]. Previous studies investigating performance indicators in soccer, however, often suffer from limitations and/or methodological problems, such as small sample sizes and univariate analyses of the observed variables [6]. If presented in isolation, a single set of indicators representing the performance of an individual or a team can give a distorted impression of a performance by ignoring other, more or less important, variables [2], which likely explains differences between previous studies [6]. It is therefore important that variables are not considered in isolation when relating performance indicators to game performance.

Technical innovations have led to an increasing availability of different performance indicators. Advanced metrics can currently be calculated in soccer using tracking data [14]. As tracking data results in millions of data points per season [15], calculating these metrics challenges data management and analytical methods of analysts [16]. Advanced metrics based on tracking data are also often provided by professional sports data companies, but obtaining these metrics is usually not free. So although metrics based on tracking data may better capture the complex nature of soccer [14], obtaining these metrics may at the moment not be feasible because of pratical and/or financial reasons for many teams. Furthermore, although there has been an increased interest into the utilization of tracking data by academics, practitioners still tend to use more traditional performance indicators [17].

In contemporary soccer, there is a wealth of different performance indicators, describing technical, tactical and physical performances, as well as contextual information. In order to provide to-the-point information to the coaching staff, a selection of the most important performance indicators can be helpful. Therefore, this study aimed to identify the strongest predictive variables of winning and losing in soccer using a broad range of variables. Machine learning is used instead of inferential statistics, as machine learning is better at handling large amounts of input variables [18]. Based on previous research, it can be hypothesized that shot-related variables such as shots and shots on target are closely related to performance in soccer; however, it is also expected that other indicators will be amongst the most important predictors.

2. Materials and Methods

2.1. Sample

All data was collected in the highest Belgian soccer division, over the course of 3 seasons (2017–2018, 2018–2019, 2019–2020), totalling 771 games. It should be noted that, because of the COVID19 pandemic, the last part of the 2019–2020 season was not played. Games with missing values because of measurements errors, such as missing physical data and/or technical data, and games that resulted in a draw, were excluded from the analysis. This resulted in the availability of 576 games for the analysis. All games were tracked using the SportVU system (Stats Perform, Chicago, IL, USA), an optical tracking system using three high-definition camera's [19]. Consent was given by STATS Perform and the Belgian Pro League for the use of the data for scientific purposes. The reliability of the data delivered by professional sports data companies, is shown to be high [20,21]. The study was conducted in accordance with the Declaration of Helsinki.

2.2. Variables

A total of 100 variables were included into the analysis (Table 1). Variables were selected based on a combination of expert knowledge and availability. All variables, with the exception of contextual game information, were derived from three sources, STATS Viewer, STATS Dynamix and STATS Edge, all offered by STATS Perform. Data regarding the teams' playing styles and Expected Goals were derived from STATS Edge. Definitions on Playing Styles and the calculation of Expected Goals can be found on the website of STATS Perform (https://www.statsperform.com, accessed on 28 June 2020). Physical game data was available at STATS Dynamix. The default speed, acceleration and deceleration thresholds for physical data were used, including the minimum effort duration of 0.5 s set for variables related to speed, accelerations and deceleration. All other variables, with the exception of contextual game information, were derived from STATS Viewer. Inside the STATS Viewer software, the user can derive its own metrics, based on criteria such as destination and direction. Direction could be set in three different directions, all relative to the opponent's goal. A ball played in an angle of $-45°$ to $+45°$

was defined as forward, backwards was defined if the ball was played in an angle from −135° to +135° relative to the opponent's goal. Balls played in an angle of +45° to +135°, and −45° to −135° were defined as sideways.

Table 1. An overview on the 100 included variables. ^{A} Playing styles: Maintenaince, Build-Up, Sustained Threat, Fast Tempo, Direct Play, Counter Attack, Crossing, High Press.

Category	**Expression**	**Parameter**	**Part Game**
Shot-related	Number (n)	- Shots, shots on target, shots not on target - Shots/shots on target/shots not on target inside attacking penalty box - Expected Goals - Shots resulting from 8 playing styles ^{A}	Full Game
Defense	Number (n)	- Duels won, possession won/loss	Full Game
Technical	Number (n)	- Total/forward/sideways/backward passes, successful passes - Total/forward/sideways/backward passes to attacking half/third - Dribbles, successful dribbles - Ball touches in attacking penalty box - Passes < 10 m, Passes < 25 m, Passes > 25 m - Possession in 8 playing styles ^{A}	Full Game
	Percentage (%)	- Total ball possession - Ball possession on defensive half/attacking half/attacking third/attacking penalty box	
Physical	Distance (m)	- Total distance - Distance between 0–6 km/h, 6–15 km/h, 15–20 km/h, 20–25 km/h, >25 km/h - Distance at accelerations/decelerations >2 ms^2/>3 ms^2	First/Second Half
	Number (n)	- Actions at speed >15 km/h/>25 km/h - Accelerations/decelerations >2 ms^2/>3 ms^2	
Disciplinal	Number (n)	- Fouls, fouls at attacking half, yellow cards, red cards, offside	Full Game
Set-Pieces	Number (n)	- Corners, penalties, free kick/throw-in on attacking third	Full Game
Contextual	Currency (€)	- Line-up/bench current estimated total transfer value - Line-up/bench estimated paid total transfer value	Full Game
	Age (years)	- Line-up/bench average age	
	Arbitrary Units (AU)	- ELO-ratings, Form	
	Number (n)	- Days between games	
	Match location	- Home/Away	

Contextual variables were obtained from other sources. ELO-ratings were included as a measure of team strength, provided by the API of http://www.clubelo.com, accessed on 1 July 2020, which is freely available. This source provides ELO-ratings for each team of over 50 (inter)national leagues, including the UEFA Europa League and Champions League, and has previously been used for scientific purposes [22,23]. Each game, both national and international, results in an exchange in ELO-points (Equation (1)), where *dr* is the difference in ELO-rating between two clubs and *R* is the result (1 for win, 0.5 for draw and 0 for loss). The exchange in ELO-points is higher for a win against a stronger team compared to a victory against an equally strong or weaker team, and vice versa for losses. This equation was derived from http://www.clubelo.com, accessed on 1 July 2020.

$$\Delta ELO = \left(R - \frac{1}{10^{(-dr/400)} + 1}\right)20 \tag{1}$$

Form was defined as the difference in a clubs' current ELO-rating compared to the ELO-rating before their previous game. Market values, ages and nationalities were obtained from https://www.transfermarkt.co.uk, accessed on 1 July 2020. These variables were previously used by [12] for scientific purposes. Lastly, the number of days between games was used as a variable, also including other competitive games such as cup and European games, to consider the possible effect of additional games for one team compared to the other on winning or losing.

Differences between the two teams competing during each game were calculated and used as input variable in this study, with the exception of match location. During games, there are interactions between the two competing teams [1], and these interactions will result in different performance statistics for both teams. The difference of each performance statistic between the two teams may provide insights into the difference in performance by the two teams on the pitch, and therefore be more informative to the model than the separate performance statistics of both the teams. Moreover, by not including the performance statistics of both teams into the model, the dimensionality of the model can be limited.

2.3. Procedures

Data from all sources was loaded into Python (version 3.7.1). To avoid multicollinearity, a Variance Inflation Factor (VIF) analysis was conducted, with a threshold of 5, using the *statsmodels* package. BorutaShap was applied as a feature selection technique, using the *BorutaShap* package. Extreme Gradient Boosting, a tree-based machine learning technique, was applied for both BorutaShap and predicting game outcome (win or lose), using the *xgboost* package (distributed by *SkLearn*). A 5-fold stratified cross-validation was used to validate the results. Cross-validation uses a large part of the data to fit the model, in this study 80%, and a small part of the data to test the model, in this study 20% [24]. Each part was thus used 4 times to fit the model and once for validation. Stratified K-fold was used to preserve balance between the frequency of each class of the dependent variable. The *StratifiedKFold* and *cross_val_score* packages, distributed by *SkLearn*, were used for the cross-validation. The average classification accuracy, precision, recall and F1-score of each cross-validation fold is reported, as well as the standard deviation.

The aim of this study was to identify the best predictive variables, therefore, it was opted to apply a tree-based machine learning, which has the advantage of high interpretability and the possibility to apply a theoretically well grounded method such as TreeExplainer to identify the strongest predictors [25]. TreeExplainer uses Shapley values to explain the global model structure, by combining local explanations of each prediction [25]. Using TreeExplainer, it is possible to determine the importance of each feature [26], on a global and local level. TreeExplainer was applied using the *shap* package. A graphical illustration on the workflow from the dataset to the identification of the best predictors is depicted in Figure 1.

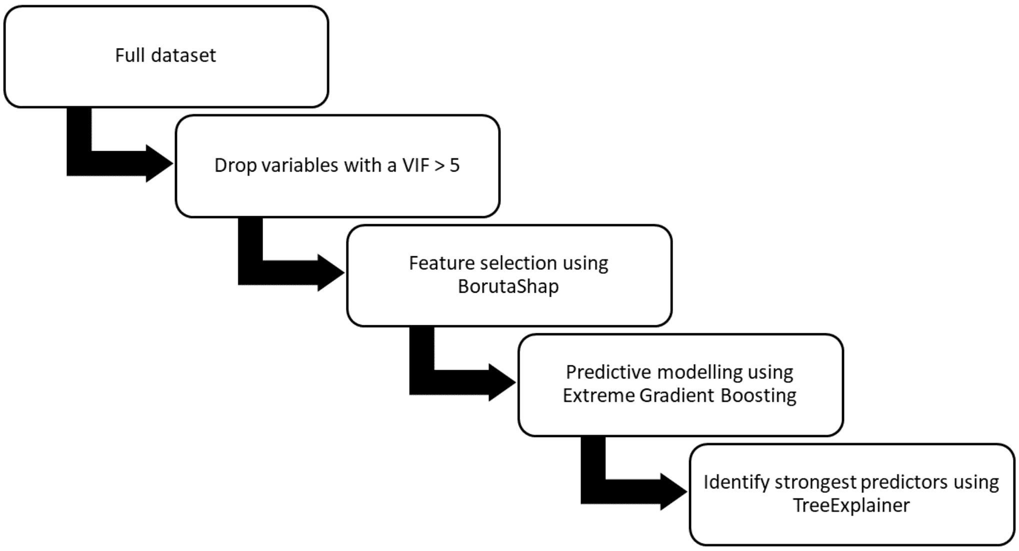

Full dataset
Drop variables with a VIF > 5
Feature selection using BorutaShap
Predictive modelling using Extreme Gradient Boosting
Identify strongest predictors using TreeExplainer

Figure 1. Graphical illustration on the workflow used for this study.

3. Results

After the removal of variables using VIF and the feature selection procedure using BorutaShap, a total of 13 variables were used during the modelling procedure. The model showed a predictive accuracy of 89.6% ± 3.1%, correctly classifying 516 out of 576 games that resulted in a win or loss (precision: 88.9%; recall: 90.1%, f1-score: 89.5%; Figure 2). In Table 2, the misclassifications are further specified in relation to total goal difference.

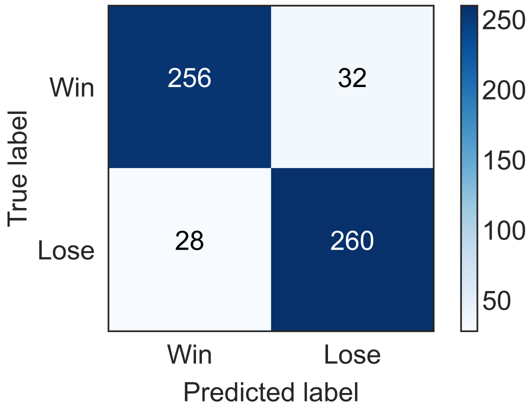


Figure 2. Confusion matrix showing the true and predicted results.

The most important predictors of the model are presented in Figure 3. As an illustration of the possibilities of local explanations using TreeExplainer, two individual game predictions are presented in Figure 4a,b.

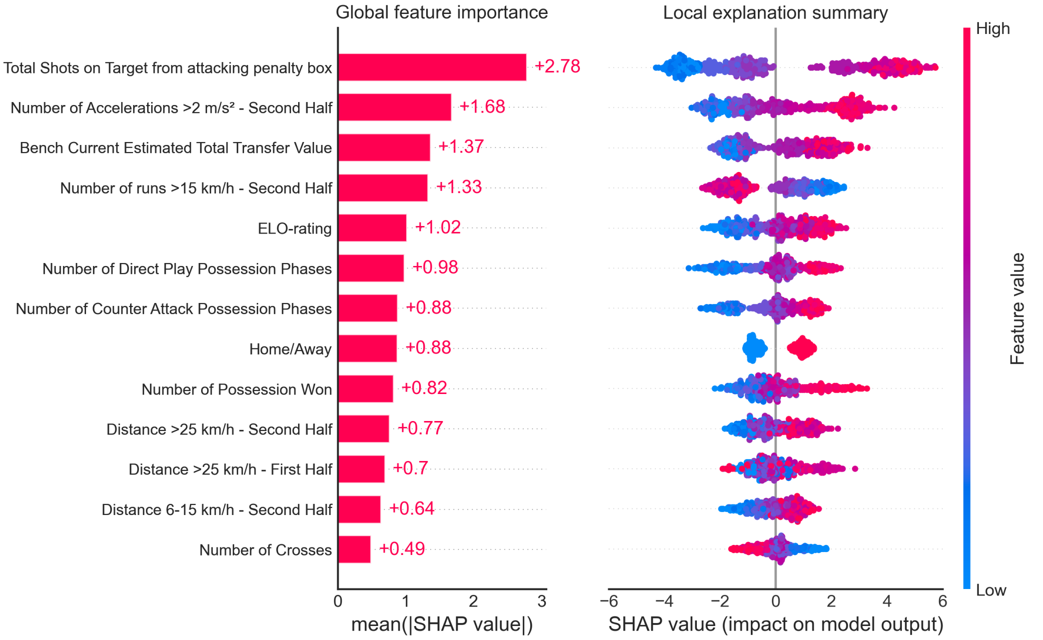
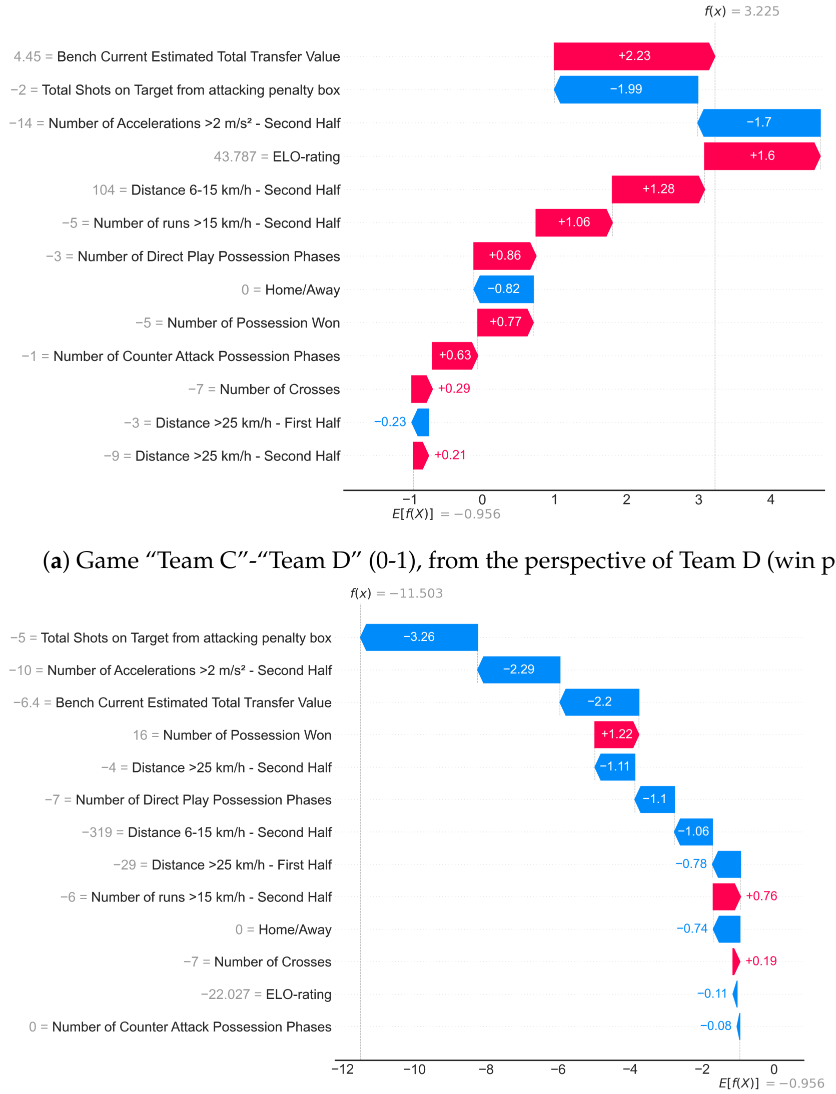


Figure 3. Feature importance based on SHAP-values. On the left side, the mean absolute SHAP-values are depicted, to illustrate global feature importance. On the right side, the local explanation summary shows the direction of the relationship between a variable and game outcome. Positive SHAP-values are indicative of winning, while negative SHAP-values are indicative of losing. As demonstrated by the colorbar, higher values are shown in red, while lower values are shown in blue. As an example, this shows that positive differences between total shots on target from the attacking penalty box between teams are associated with winning, while negative differences are associated with losing.

Table 2. An overview on the total and relative number of games and misclassifications for each goal difference.

Goal Difference	Games (n = 576)		Misclassifications (n = 60)	
	Total	**%**	**Total**	**%**
1	302	52.4%	43	71.7%
2	160	27.8%	16	26.7%
3	69	12.0%	1	1.7%
4	33	5.7%	0	0%
5	8	1.4%	0	0%
6	3	0.5%	0	0%
7	1	0.2%	0	0%



(**a**) Game "Team C"-"Team D" (0-1), from the perspective of Team D (win predicted).

(**b**) Game "Team A"-"Team B" (2-1), from the perspective of Team B (loss predicted).

Figure 4. Two waterfall plots exampling local game predictions, showing the contribution of each variable to the prediction. The variable name is preceded by the value of the particular variable. Below the x-axis, the baseline value (E[f(X)]) is displayed, indicative of the expected value of the model evaluated on the background dataset. The SHAP-values of each variable are summed to match the model output with all variables included. Positive SHAP-values push the model to predict a win, while negative SHAP-values push the model to predict a loss.

4. Discussion

This study aimed to identify the strongest predictive variables of winning and losing in Belgian professional soccer. A broad spectrum of variables was used to build a predictive model. The results showed that more than 89% of the games resulting in a win or loss could be correctly classified. Total shots on target from the attacking penalty box showed to be the strongest predictor. Interestingly, a broad range of variables, including physical indicators such as the distance in several speed zones, the number of accelerations (>2 m/s^2) and number of actions >15 km/h showed to be among the most important variables related to game results, as well as contextual variables such as ELO-rating, the total added transfer value of the benched players and match location.

There are different purposes for predicting game outcome. Previous studies relating to the prediction of game outcome in soccer often focused on betting [27]. These studies used historic games to construct a model, which was then evaluated by predicting future soccer games. In our study however, the aim was to identify the strongest predictive variables, so instead of a division between historic and future games, a cross-validation approach was used to evaluate model performance. The prediction accuracy was considerably higher compared to a similar study conducted by [28], who reported classification accuracies of 72.7% and 83.3% when predicting respectively losing and winning in professional soccer using artificial neural networks. A classification accuracy of amply 89% can be considered as high, as it has been shown that chance plays a major role in goal-scoring [29]. The results show that the majority of the misclassifications occur when the final goal difference between two teams is small. Given these small goal differences between teams, and the impact that each goal has on game outcome [29], the occurrence of "lucky winners" or "unlucky losers" is frequent in low-scoring sports [30], and classification accuracies of close to 100% seem improbable.

It was hypothesized that shot-related variables were among the strongest predictors, as previous studies already showed that shot-related variables closely relate to game outcome in soccer [6,8,10,12]. In this study, the total shots on target from the attacking penalty box showed to be the best predictor of winning and losing. Of all shot-related variables, total shots on target from the attacking penalty box was the only variable which was not rejected by VIF or BorutaShap, showing that shot-related variables are closely interrelated. It should be noted that this does not directly indicate that other shot-related variables can be deemed as unimportant, but that the information entailed by those measures is already captured, or better captured, by other metrics in relation to game outcome. The number of shots on or off target are often reported by sports data providers, as opposed to location, which is usually not reported. It may therefore be useful to either use both total shots on target from the penalty box, or use a metric such as Expected Goals, which also takes shot location into account. Other often-reported metrics such as the total number of passes and the number of successful passes are not among the best important predictors of game outcome, which may also be explained by the informative value to the model in comparison to other variables. This may also explain why Playing Styles-related possessions, such as Direct Play and Counter Attack, are amongst the best predictors, as they may provide more information to the model than total ball possession as an isolated metric.

Physical fitness is deemed as an important factor relating to performance in soccer, however, the role of physical game output in relation to other performance indicators remained to be elucidated [31]. In our study, several physical indicators are shown to be among the best predictors of game outcome in soccer. All variables relating to physical variables were subdivided into the first and second half, as it was previously shown that physical game performances, such as high-speed running [20], the number of acc- and decelerations and the distance in several acc- and deceleration zones decrease at the end of the game [32]. Interestingly, most of the selected predictive physical variables relate to the second half, with the exception of the distance >25 km/h, of which both halves were included into the modelling procedure. This study also shows that total difference in the number of medium accelerations (>2 m/s^2) is associated with winning or losing, further confirming the importance of high-intensity efforts in soccer [33]. As indicated by the inclusion of the variable distance between 6–15 km/h in the second half, the ability to maintain the physical capabilities to not only perform high intensity actions, but also low to medium intensity efforts throughout the game seems to be important. With regard to the interpretation of physical performance indicators, it is however important to note that "more" does not always indicate "better", as shown by the inverse relationship of the number of actions >15 km/h and game outcome. This finding is also partly confirmed by the study [34], showing that players covered less distance in the zone between 17–21 km/h during games that were won. As physical game output depends on a myriad of factors, such as ball possession [35], pacing strategy [36], match location and match status [37], physical game output, regardless of its expression, should be viewed in relation to other performance indicators [31,38] and contextual information [37].

Some variables that can be related to attacking play are negatively associated to game outcome. In accordance with previous research [39,40], higher frequencies of crosses are negatively associated with game outcome. Crossing, which can be defined as an airborne delivery of the ball into the opponent's penalty area [41], may therefore be labelled as an inefficient method to create good scoring opportunities [40,41]. It should, however, be noted that playing style depend on the qualities and characteristics of the team [11]. Therefore, the coaching staff may decide to apply a playing style that can generally be characterized as inefficient, because it matches the teams' qualities and characteristics.

In a low-scoring game such as soccer, rare events are often those that lead to success [14]. These events should be captured to accurately predict game outcome, however, these events cannot always be properly quantified, even with more advanced metrics based on tracking data. Actions that are currently difficult to quantify, such as good positioning or the ability to give defense-splitting passes, are however often recognized by clubs, media and/or fans, resulting in higher transfer values reported by sources such as https://www.transfermarkt.co.uk, accessed on 5 March 2021. These actions also help teams to get better game outcomes, resulting in improved ELO-ratings. Including transfer values and ELO-ratings may thus be useful, possibly by partly filling the gap of what cannot be (currently) quantified, also considering the feasibility of obtaining these variables in terms of practical and financial reasons.

Machine learning was used in this study to identify the strongest predictive variables. It has also previously been used in a soccer context not only in relation to game outcome [42–44] and tactics [45], but also in relation to training load [46,47] and injuries [48], showing the broad window of applications of machine learning in soccer. Given that game performance [49], training load [50] and injury [51] are all multidimensional, the application of machine learning can be useful since it is particularly helpful when dealing with many input variables. Developments in the area of machine learning, such as TreeExplainer [25], which is not only a strong theoretically grounded method to calculate feature importance [26], but also allows to build visualisations that indicate the direction of the relation of a variable with performance, can be helpful in the translation from science to practice. Illustrations such as those displayed in Figures 3 and 4 can be useful for analysts to show how features impact game outcome.

Future studies should attempt to add more detailed information for several features, for example, total distance in and out of possession or detailed information on the position of crosses. This information can be informative to the model and aid the explanation of results. As data is often provided by sports data companies, these companies should be encouraged to add more detail to the provided data to allow more thorough analyses. The use of "new" features should also be encouraged, to test their added value in relation to other, more established features. It should also be noted that the results from this study cannot fully distinguish whether a variable is the cause of a (un)favourable scoreline, or the effect. To illustrate, losing teams attempt to turn the game and therefore may engage in more high-intensity efforts [34]. The winning side on the other hand, may fall back, allowing less space for losing side to play and perform sprints, while the losing team may apply a more risk-taking strategy, which could result in more counter-attacking play of the winning side [39,52]. Therefore, more research is necessary to gain more insight into the cause-effect relation between performance indicators and game outcome.

The results from our study show which variables can be considered as the best predictors of an accurate model predicting winning and losing in professional Belgian soccer. It provides the direction of the relationship of these variables with winning and losing, also in relation to the other predictors. It was shown that not only shot-related variables, but a broad range of variables are amongst the strongest predictors of winning and losing. As the workflow from dataset to predictive modelling was also described in detail, similar approaches can be used to evaluate the current performance indicators provided to the coaching staff and other stakeholders connected to the team. It seems particularly interesting to look at physical parameters of the second half, given that they are amongst the best predictors of game outcome in soccer. Variables such as ELO-ratings, transfer values, match location and Playing Styles can be useful additions to current approaches used to evaluate game performances.

Author Contributions: Conceptualization, Y.G., J.B., S.V. and J.G.B.; methodology, Y.G., J.B., S.V. and J.G.B.; software, Y.G., and S.V.; validation, Y.G., J.B., S.V. and J.G.B.; formal analysis, Y.G., and S.V.; investigation, Y.G..; resources, Y.G.; data curation, Y.G. and S.V.; writing—original draft preparation, Y.G.; writing—review and editing, Y.G., J.B., S.V. and J.G.B.; visualization, Y.G., and S.V.; supervision, Y.G., J.B., S.V. and J.G.B.; project administration, Y.G. and J.G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Part of the data was obtained from STATS Perform. Please contact Jan G. Bourgois (jan.bourgois@ugent.be) to inform about the data availability.

Acknowledgments: The authors would like to thank team of Steven Verstockt.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VIF	Variance Inflation Factor

References

1. Lames, M.; McGarry, T. On the search for reliable performance indicators in game sports. *Int. J. Perform. Anal. Sport* **2017**, *7*, 62–79. [CrossRef]
2. Hughes, M.D.; Bartlett, R.M. The use of performance indicators in performance analysis. *J. Sports Sci.* **2002**, *20*, 739–754. [CrossRef]
3. Plessner, H.; Haar, T. Sports performance judgments from a social cognitive perspective. *Psychol. Sport Exerc.* **2006**, *7*, 555–575. [CrossRef]
4. Mark, M.M.; Mutrie, N.; Brooks, D.R.; Harris, D.V. Causal Attributions of Winners and Losers in Individual Competitive Sports: Toward a Reformulation of the Self-Serving Bias. *J. Sport Psychol.* **1984**, *6*, 184–196. [CrossRef]
5. Brechot, M.; Flepp, R. Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *J. Sports Econ.* **2020**, *21*, 335–362. [CrossRef]
6. Castellano, J.; Casamichana, D.; Lago, C. The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. *J. Hum. Kinet.* **2012**, *31*, 139–147. [CrossRef] [PubMed]
7. Collet, C. The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010. *J. Sports Sci.* **2013**, *31*, 123–136. [CrossRef] [PubMed]
8. Lago-ballesteros, J.; Lago-Peñas, C. Performance in Team Sports: Identifying the Keys to Success in Soccer. *J. Hum. Kinet.* **2010**, *25*, 85–91. [CrossRef]

9. Yang, G.; Leicht, A.S.; Lago, C.; Gómez, M.-Á. Key team physical and technical performance indicators indicative of team quality in the soccer Chinese super league. *Res. Sports Med.* **2018**, *26*, 158–167. [CrossRef] [PubMed]
10. Broich, H.; Mester, J.; Seifriz, F.; Yue, Z. Statistical Analysis for the First Bundesliga in the Current Soccer Season. *Prog. Appl. Math.* **2011**, *7*, 1–8.
11. Harrop, K.; Nevill, A. Performance indicators that predict success in an English Professional League One Soccer Team. *Int. J. Perform. Anal. Sport* **2014**, *14*, 907–920. [CrossRef]
12. Lepschy, H.; Wäsche, H.; Woll, A. Success factors in football: An analysis of the German Bundesliga. *Int. J. Perform. Anal. Sport* **2020**, *20*, 150–164. [CrossRef]
13. Ruiz-Ruiz, C.; Fradua, L.; Fernández-García, Á.; Zubillaga, A. Analysis of entries into the penalty area as a performance indicator in soccer. *Eur. J. Sport Sci.* **2013**, *13*, 241–248. [CrossRef] [PubMed]
14. Goes, F.R.; Kempe, M.; Lemmink, K. Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data. In *Mathsport International Conference Proceeding*; Propobos Publications: Athens, Greece, June 2019; pp. 105–115.
15. Bialkowski, A.; Lucey, P.; Carr, P.; Yue, Y.; Sridharan, S.; Matthews, I. Large-scale analysis of soccer matches using spatiotemporal tracking data. In Proceedings of the IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014; pp. 725–730.
16. Goes, F.R.; Meerhoff, R.L.A.; Bueno, M.J.; Rodrigues, D.M.; Moura, F.A.; Brink, M.S.; Elferink-Gemser, M.T.; Knobbe, A.J.; Cunha, S.A.; Lemmink, K.A. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *Eur. J. Sport Sci.* **2020**, 1–16. [CrossRef] [PubMed]
17. Perin, C.; Vuillemot, R.; Stolper, C.D.; Stasko, J.T.; Wood, J.; Carpendale, S. State of the Art of Sports Data Visualization. *Comput. Graph. Forum* **2018**, *37*, 663–686. [CrossRef]
18. Bzdok, D.; Altman, N.; Krzywinski, M. Points of Significance: Statistics versus machine learning. *Nat. Methods* **2018**, *15*, 233–234. [CrossRef] [PubMed]
19. Linke, D.; Link, D.; Lames, M. Validation of electronic performance and tracking systems EPTS under field conditions. *PLoS ONE* **2018**, *13*, e0199519. [CrossRef] [PubMed]
20. Bradley, P.; Di Mascio, M.; Peart, D.; Olsen, P.; Sheldon, B. High-Intensity Activity Profiles of Elite Soccer Players at Different Performance Levels. *J. Strength Cond. Res.* **2010**, *24*, 2343–2351. [CrossRef]
21. Bradley, P.; O'Donoghue, P.; Wooster, B.; Tordoff, P. The reliability of ProZone MatchViewer: A videobased technical performance analysis system. *Int. J. Perform. Anal. Sport* **2007**, *7*, 117–129. [CrossRef]
22. Csató, L. The UEFA Champions League seeding is not strategy-proof since the 2015/16 season. *Ann. Oper. Res.* **2020**, *292*, 161–169. [CrossRef]
23. Engist, O.; Merkus, E.; Schafmeister, F. The Effect of Seeding on Tournament Outcomes: Evidence From a Regression-Discontinuity Design. *J. Sports Econ.* **2021**, *22*, 115–136. [CrossRef]
24. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Number 2; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009.
25. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
26. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 15 November 2020).
27. Dubitzky, W.; Lopes, P.; Davis, J.; Berrar, D. The Open International Soccer Database for machine learning. *Mach. Learn.* **2019**, *108*, 9–28. [CrossRef]
28. Hassan, A.; Akl, A.R.; Hassan, I.; Sunderl, C. Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. *Sensors* **2020**, *20*, 3213. [CrossRef] [PubMed]
29. Lames, M. Chance involvement in goal scoring in football—An empirical approach. *Ger. J. Exerc. Sport Res.* **2018**, *48*, 278–286. [CrossRef]
30. Simon, R. Deserving to be lucky: Reflections on the role of luck and desert in sports. *J. Philos. Sport* **2007**, *34*, 13–25. [CrossRef]
31. Carling, C. Interpreting physical performance in professional soccer match-play: Should we be more pragmatic in our approach? *Sports Med.* **2013**, *43*, 655–663. [CrossRef]
32. Russell, M.; Sparkes, W.; Northeast, J.; Cook, C.J.; Love, T.D.; Bracken, R.M.; Kilduff, L.P. Changes in Acceleration and Deceleration Capacity Throughout Professional Soccer Match-Play. *J. Strength Cond. Res.* **2016**, *30*, 2839–2844. [CrossRef] [PubMed]
33. Faude, O.; Koch, T.; Meyer, T. Straight sprinting is the most frequent action in goal situations in professional football. *J. Sports Sci.* **2012**, *30*, 625–631. [CrossRef] [PubMed]
34. Chmura, P.; Konefał, M.; Chmura, J.; Kowalczuk, E.; Zajac, T.; Rokita, A.; Andrzejewski, M. Match outcome and running performance in different intensity ranges among elite soccer players. *Biol. Sport* **2018**, *35*, 197–203. [CrossRef]
35. Dellal, A.; Chamari, K.; Wong, D.P.; Ahmaidi, S.; Keller, D.; Barros, R.; Bisciotti, G.N.; Carling, C. Comparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga. *Eur. J. Sport Sci.* **2011**, *11*, 51–59. [CrossRef]
36. Paul, D.J.; Bradley, P.S.; Nassis, G.P. Factors affecting match running performance of elite soccer players: Shedding some light on the complexity. *Int. J. Sports Physiol. Perform.* **2015**, *10*, 516–519. [CrossRef] [PubMed]

37. Lago, C.; Casais, L.; Dominguez, E.; Sampaio, J. The effects of situational variables on distance covered at various speeds in elite soccer. *Eur. J. Sport Sci.* **2010**, *10*, 103–109. [CrossRef]
38. Bradley, P.; Ade, J.D. Are Current Physical Match Performance Metrics in Elite Soccer Fit for Purpose or is the Adoption of an Integrated Approach Needed? *Int. J. Sports Physiol. Perform.* **2018**, *13*, 656–664. [CrossRef] [PubMed]
39. Fernandez-Navarro, J.; Fradua, L.; Zubillaga, A.; McRobert, A.P. Influence of contextual variables on styles of play in soccer. *Int. J. Perform. Anal. Sport* **2018**, *18*, 423–436. [CrossRef]
40. Liu, H.; Gomez, M.Á.; Lago-Peñas, C.; Sampaio, J. Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *J. Sports Sci.* **2015**, *33*, 1205–1213. [CrossRef] [PubMed]
41. Vecer, J. Crossing in Soccer has a Strong Negative Impact on Scoring: Evidence from the English Premier League the German Bundesliga and the World Cup 2014. Technical Report. 30 September 2014. Available online: https://ssrn.com/abstract=2225728 (accessed on 5 March 2021).
42. Constantinou, A.C. Dolores: A model that predicts football match outcomes from all over the world. *Mach. Learn.* **2019**, *108*, 49–75. [CrossRef]
43. Hucaljuk, J.; Rakipović, A. Predicting football scores using machine learning techniques. In Proceedings of the 2011-34th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, Croatia, 23–27 May 2011; Volume 48, pp. 1623–1627.
44. Stübinger, J.; Mangold, B.; Knoll, J. Machine learning in football betting: Prediction of match results based on player characteristics. *Appl. Sci.* **2020**, *10*, 46. [CrossRef]
45. Memmert, D.; Lemmink, K.A.; Sampaio, J. Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sports Med.* **2017**, *47*, 1–10. [CrossRef]
46. Geurkink, Y.; Vandewiele, G.; Lievens, M.; De Turck, F.; Ongenae, F.; Matthys, S.P.; Boone, J.; Bourgois, J.G. Modeling the Prediction of the Session Rating of Perceived Exertion in Soccer: Unraveling the Puzzle of Predictive Indicators. *Int. J. Sports Physiol. Perform.* **2018**, *14*, 1–6. [CrossRef]
47. Jaspers, A.; Brink, M.S.; Probst, S.G.; Frencken, W.G.; Helsen, W.F. Relationships Between Training Load Indicators and Training Outcomes in Professional Soccer. *Sports Med.* **2017**, *47*, 533–544. [CrossRef] [PubMed]
48. Rommers, N.; Rössler, R.; Verhagen, E.; Vandecasteele, F.; Verstockt, S.; Vaeyens, R.; Lenoir, M.; D'Hondt, E.; Witvrouw, E. A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Med. Sci. Sports Exerc.* **2020**, *52*, 1745–1751. [CrossRef] [PubMed]
49. Stolen, T.; Chamari, K.; Castagna, C.; Wisloff, U. Physiology of Soccer. *Sports Med.* **2005**, *35*, 501–536. [CrossRef] [PubMed]
50. Impellizzeri, F.M.; Rampinini, E.; Marcora, S.M. Physiological assessment of aerobic training in soccer. *J. Sports Sci.* **2005**, *23*, 583–592. [CrossRef] [PubMed]
51. Bahr, R.; Krosshaug, T. Understanding injury mechanisms: A key component of preventing injuries in sport. *Br. J. Sports Med.* **2005**, *39*, 324–329. [CrossRef] [PubMed]
52. Lago, C. The influence of match location, quality of opposition, and match status on possession strategies in professional association football. *J. Sports Sci.* **2009**, *27*, 1463–1469. [CrossRef]