Soccer Events Analysis

R. Manetti, G. Martini, N. Ozgoli, and M. Razzaghnoori

University Of Pisa

Abstract. In recent times, the advent of big data analysis techniques in soccer became very popular. In this scenario, the present work has the purpose of improving the state-of-the-art results in the very hard task of making predictions about the goals scored by a team.

Introduction. When dealing with a multitude of annotated events taking place during a football match, a wise usage of the information available is crucial for providing predictions about the outcome of a match. Such predictions can be used to allow a more "efficient" usage of the player's talent, preserving their health together with improving the chances of winning a competition.

The remaining part of this report is structured in three main parts. In Section 1 we expand the raw data we used as an input, and describe the main modifications we performed on them and the results obtained. In Section 2 we describe the analysis done for the selection of the classifier and the evaluation of them with some baselines. Finally in Section 5 we make our consideration of the entire analysis.

1 Data understanding

1.1 (not) The data we deserve

The dataset[1] we've been working on contains information about some competitions (from Serie A, to Bundesliga, to World Cup) that took place during an entire season of matches.
Such dataset is obtained from Wyscout and contains many files, each of the ones concerns a certain field, as follows:

- Coaches: where all the information about the coaches is stored;
- Referees: containing some info about the referees;
- Players: which stores information about the players involved;
- Teams: the collection where the information about the teams is stored;
- Competitions: contains a lot of details about the competitions under consideration;
- Events: is the collection more significant and contains all the meaningful events that took places in every match in the dataset. Apart from the insider-only fields (aka identifiers), the most relevant fields contained in this collection are the eventName and tag;
- Matches: which stores information about all the matches.

[1] Such a dataset is public and can be found here.

- Playerank[2]: is an extra collection related on the previous data because it contains the playerank score for each player that have played in a match.

It goes without saying that this multitude of data need to be refined in order to obtain useful insights about a match, moreover a very demanding research on the semantics of data was held so that one may have a full understanding of the variables available.

To fulfill this purpose, a process of data transformation was performed, the reader can find the results in Section 1.2.
In the dataset are 7 different competitions, 2 of them are ”national” and the remaining 5 are ”club”. The the number of events and match that we have analyzed are resumed below.

	IT	ES	EN	DE	FR	WC	EU	total
events	647372	628659	643150	519407	632807	101759	78140	3251294
matches	380	380	380	306	380	64	51	1941

1.2 (but) The data we need

Before digging into the details of the variables that have been created, it is important to spend a few lines introducing the predictive task we are focusing on.
We designed three different targets for this analysis, based on the behaviour of the team during a time window, that for now is the first half of a game. The first is (1) how likely it is for a team to win (isWinner), then (2) how likely it is for the same team to score a goal in the time window of prediction (didScoreInTWP), that for now corresponds to the second half of the match, and finally (3) the goals difference between the goals scored by team actually analyzed and the goals scored by the opponent team (goalsDiff).
According to common sense reasoning and after reading some soccer-related commentary, the authors decided to transform the json-like dataset into a table, where the match (studied from the point of view of a certain team) is uniquely identified by the couple (matchId, teamId) and the features follow[3]:

- isHome: 1 if the team played the match in it’s home city, 0 otherwise.
- teamPoints: points scored by the team in the competition until now.
- teamPlayerank: for each player in the team, the exponentially weighted moving average of his playerank scores obtained in previous played matches is computed, then the mean of all these values is evaluated.
- meanPlayerOverall and meanPlayerPotential: represent the “goodness” of a team (measured as the average of the “goodness” on all the players belonging to that team) and the average potential of growth among all the players of that team respectively.

[2] To learn how the rank of the players is evaluated you can click here.

[3] The first two variables are obtained joining the Wyscout dataset with another dataset containing the information of the players of Fifa 2019 competition.

- meanPrevScore, meanPrevScoreET, meanPrevScoreHT and meanPrevScoreP: correspond to the average number of goals scored in previous matches divided for periods: all the match, the extra time (ET), the first half of the match (HT) and penalties (notice that exra time and penalties are assigned only in international competitions).
- numDuel, numFoul, numFreeKick, numGoalkeeperLeavingLine, numInterruption, numOffside, numOthersOnTheBall, numPass, numSaveAttempt and numShot: correspond to the sum of events done by a team in a match during the first half of the match.
- rateAccFreeKick, rateAccPass and rateAccShot: are respectively rate of accurate free kicks, passes and shots performed.
- numRedCard, numYellowCard and numSecondYellowCard: are respectively the number of red, yellow and second yellow (that lead to expulsion) cards assigned to the team during the first half of a match.
- percBallPoss adn percOppHalfField: are the percentage of ball possession of the team, respectively on all the events and on which are done in the opponent's side of field, in the time window analyzed.
- numGoalsTW, numOwnGoalsTW, numGoalsTWP and numOwnGoalsTWP: are respectively the number of goals and own goals done by the team in the time window analyzed (TW) and of prediction (TWP).
- scoreTW: is the score of the team at the end of the time window analyzed, that corresponds to the sum of the field numGoalsTW of that team and the field numOwnGoalsTW of the opponent team.
- numGoals1H, numGoals2H and numGoalsET: are the number of goals done respectively in the first, second half of the match and extra time.
- finalScore: is the score of the team at the end of the match.

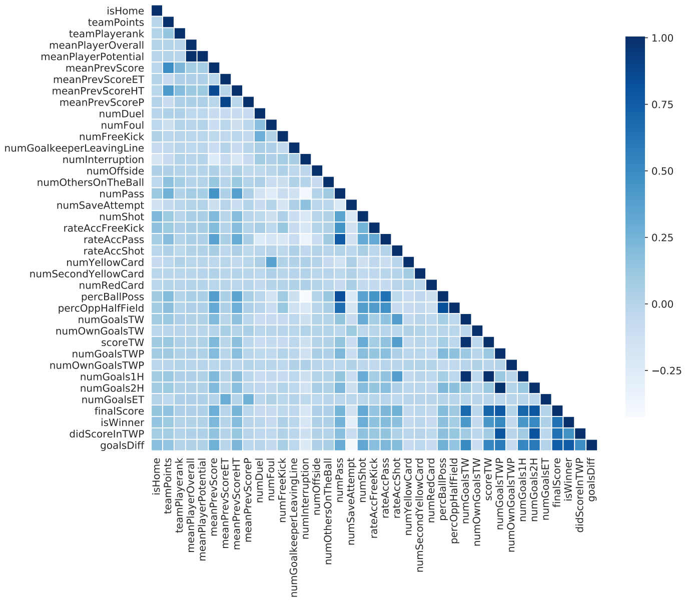

Fig. 1. Correlation matrix

1.3 Experimental results

After the extrapolation of the data and creation of some features, a phase of feature reduction took place, with the aim of better study all the events that happened during the first half time of each match and each team.
The correlation matrix can be found in Figure 1, where the last three rows/columns are the target variables. An attentive reader may notice that there are only a few features that are very well correlated. The number of passes and the rate of the accurate ones are two variables which are well correlated, but we observed that in any match and and for any team there are approximately the 80% of accurate passes (see Figure 2). Moreover, the features meanPlayerOverall and meanPlayerPotential are well correlated, but in this case the distribution of both are very similar. Finally, we state that the fields percBallPoss and percOppHalfField are correlated with the numPass and numAccuratePass (and also with rateAccuratePass, following the same reasoning described above).
In order to provide a statistical analysis of the features we designed there is Figure 2, where all the 3 million events are taken into account to compute the average number.

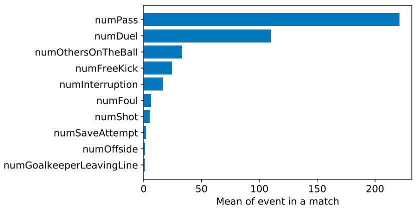

Fig. 2. Mean of each event happened during one match.

On the other hand, the number of goals scored during the first half have the behaviour shown in the rest of the section.
Figure 3 represents the histogram of the two target variables under consideration. The plot on the left, which is based on the target isWinner shows clearly that if in the first half of a match a certain team has not scored any goal, the winning chances for such team are very low. "The biggest the number of goals scored during the first half, the more likely it is for such team to win", which holds except for the case in which the number of goals is 1 (and where the probability to win or lose is approximately the same).
Conversely, for the second target variable, didScoreInTWP (right-hand plot) there is not a clear difference in probability to say if a team is likely to score a goal in the second half only on the base of the number of goals scored in the first half.
Figure 4 represents the density of the meanPrevScore, that we have evaluated on the concept that we maybe can predict the variable targets taking into account the average score of a team in the previous match.

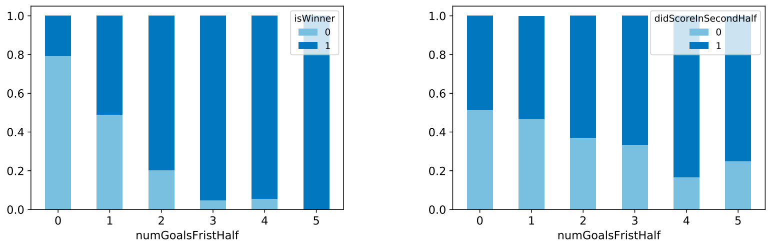

Fig. 3. Histogram of the field numGoalsFristHalf divided for the two targets.

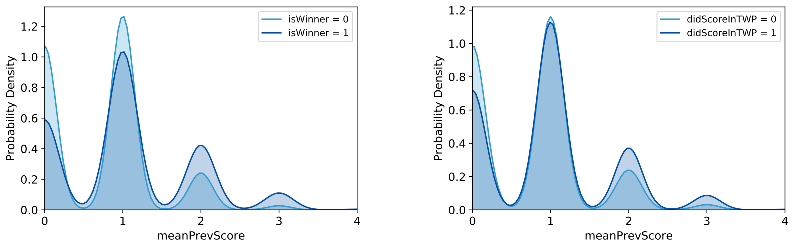

Fig. 4. Density of the field meanPrevScore divided for the two target.

It is crucial to observe the importance of the feature which models the "percentage of events happened in the opposite half field of each team". In other words, whenever a team is responsible form many events in the half field of the opponent such team has a reasonable advantage with respect to the other team.

To motivate this reasoning, a histogram based on the two targets is shown in Figure 6. It goes without saying that in order to analyze such polt it is important to take into account the distribution of the values of such a feature (see Figure 5).

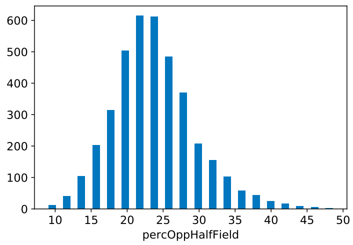

Fig. 5. Distribution of the field percOppHalfField.

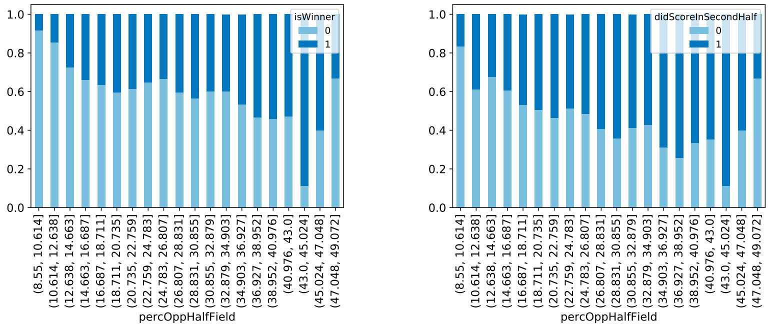

Fig. 6. Histogram of the field percOppHalfField divided for the two targets.

2 Models evaluations

This section will describe the approach used to select and the baselines implemented to evaluate the models for this project.

2.1 Approach to models selection

The training and the testing of a model implicate the subdivision of the dataset given by the previous analysis. This division is usually made by a function to which we need to provide only the dataset and its percentage that we want to use for the test. In this project exists a temporal problem that doesn't allow to use directly this kind of function. More precisely, some fields in this analysis are based on facts that happened in the past, e.g. teamPlayerank, which consider all the matches that each player of the team have played before this match, and if it is the first match of the competition played by this team, the information of the past matches doesn't exists. That can produce some errors in the test phase so we decided to put directly in the training set all the information for a specific (matchId, teamId) for which we don't have the knowledge of the past. Then the remaining part of the dataset has been divided to have at the end the 20% for tests set, while the other 80% has been added to the training set.

To find the most efficient model for our analysis we have used the GridSearchCV with hyperparameter tuning. The four classifiers evaluated are: Decision Tree (DTC), KNeighbors (KNC), Multi Lyer Perceptron (MLPC), and Support Vector (SVC). The parameters selected for each classifier are reported in Table 1.

Moreover, for each classifier we have computed the grid search for four different scoring metric: accuracy (ACC), F1, precision (PPV) and recall (TPR). For the last three scoring metrics we have used the default binary average for binary the target isWinner and didScoreInTWP, and the macro average for the multi-class target goalsDiff. All these metrics are evaluated on the predicted target with respect to the test target and with respect to some baseline, which are described in the following section.

Classifier	Parameters
DTC	criterion: [gini, entropy], max_depth: [5, 10, 15, 25, 50, 75, 100, None], splitter: [best, random], min_samples_split: [2, 5, 10, 15, 30, 50], max_features: [2, 4, 6, 8, 10, sqrt, log2, None]
KNC	weights: [uniform, distance], metric: [euclidean, manhattan], n_neighbors: [3,9,15,19]
MLPC	activation: [relu, logistic, tanh], hidden_layer_sizes: [(5,2), (10,5)]
SVC	kernel: [rbf], gamma: [1e-3, 1e-4], C: [0.1, 1, 10, 100, 1000] or kernel: [linear], C: [0.1, 1, 10, 100, 1000]

Table 1. Hyperparameter tuning for GridSearchCV.

2.2 Baselines for models evaluation

Instead of using a dummy target to evaluate the target predicted by one of the selected models, we have evaluate some baselines, one or more for each target, which are described below.

(1) For the first target, isWinner, we have a baseline
  11 to say if a team wins a match based on its goodness.
  12 to say if a team A wins a match against a team B more/less powerful, analyzing all the matches in which A won against a team more/less powerful.
(2) For the second target, didScoreInTWP, we have a baseline
  21 to say if a team did a score in the time window of prediction analyzing the same time window of all the matches available.
(3) For the third target, goalsDiff, we have a baseline
  31 to compute the difference between the number of goals scored and conceded for a team A that play a match against a team B more/less powerful, analyzing all the matches in which A have played against a team more/less powerful before the date of the actual match.
  32 to compute the difference between the number of goals scored and conceded for a team A that plays a match against a team B, analyzing all the matches played by A before the date of the actual match.

	Scoring	Accuracy				F1			
		DTC	KNC	MLPC	SVC	DTC	KNC	MLPC	SVC
test	ACC	0.73	0.73	0.75	0.75	0.72	0.73	0.74	0.75
	F1	0.55	0.55	0.62	0.75	0.55	0.55	0.64	0.62
	PPV	0.69	0.67	0.68	0.69	0.66	0.67	0.64	0.70
	TPR	0.46	0.47	0.57	0.52	0.48	0.47	0.63	0.55
baseline 11	ACC	0.76	0.75	0.70	0.73	0.64	0.75	0.65	0.72
	F1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	PPV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	TPR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
baseline 12	ACC	0.68	0.71	0.68	0.69	0.70	0.71	0.67	0.69
	F1	0.33	0.41	0.41	0.39	0.39	0.41	0.44	0.40
	PPV	0.33	0.40	0.36	0.37	0.38	0.40	0.37	0.37
	TPR	0.33	0.42	0.46	0.42	0.41	0.42	0.54	0.44

Table 2. Scoring results for target isWinner.

2.3 Resulting scores

For the analysis of the results we consider only the scoring values obtained from the execution of the GridSearchCV with accuracy and F1 scoring values. That because, according to the aim of this project and a possible scenario of usage of that analysis, we believe that these two scoring metrics are more important.

Table 2 reports, for the target isWinner, the values of the four metrics given by the evaluation of a classifier according to the test and the baselines. The execution for both scoring values, accuracy and F1, don't produce very different results. What you can say observing them is that all the classifiers are very similar, so any classifier is more efficient than the others.

Moreover, for both cases, if you pay attention for the accuracy scores gives from the evaluation of the test and the baselines, you can see that the test results are more similar to the results of baseline 11 than of baseline 12. The results of the other score metrics for both the baselines are similar in accuracy and F1. Finally the information given from the F1, precision and recall results of baseline 11 are not very useful.

Concentrating now on the analysis done for the target didScoreInTWP, which results are reported in Table 3, we can say that, for the test, each score result for each classifier is the same (or very similar) if we use the accuracy or the F1 as scoring. You can also see that the SVC for F1 gives the maximum recall and the minimum accuracy and precision, respect the other classifiers, which do not differ much from each other.

Also in this case the information given from the F1, precision and recall results of baseline 21 are not very useful. While now, we can notice a discrepancy between the accuracy results of test and baseline, the results of the last one are very low compared to the result of the test.

Observe the analysis done for the target goalsDiff, which results are reported in Table 4. Again, the results obtained from the execution of all the classifiers with accuracy and F1 scoring metric, are not so distinct, more precisely, in some case they have the same value.

What it's possible to notice is that both the baselines, which have approximately the same results, are very low than the results of the test on all the score metrics. The accuracy results of baselines are roughly half

	Scoring	Accuracy				F1			
		DTC	KNC	MLPC	SVC	DTC	KNC	MLPC	SVC
test	ACC	0.60	0.57	0.60	0.61	0.60	0.57	0.61	0.55
	F1	0.67	0.61	0.66	0.64	0.67	0.61	0.63	0.71
	PPV	0.61	0.61	0.62	0.65	0.61	0.61	0.65	0.55
	TPR	0.74	0.61	0.70	0.62	0.74	0.61	0.61	1.00
baseline 21	ACC	0.34	0.45	0.38	0.47	0.33	0.45	0.49	0.00
	F1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	PPV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	TPR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3. Scoring results for target didScoreInTWP.

the results of the test accuracy. While for the other metric results are definitely low.

	Scoring	Accuracy				F1 (macro)			
		DTC	KNC	MLPC	SVC	DTC	KNC	MLPC	SVC
test	ACC	0.60	0.57	0.60	0.61	0.60	0.57	0.61	0.55
	F1	0.67	0.61	0.66	0.64	0.67	0.61	0.63	0.71
	PPV	0.61	0.61	0.62	0.65	0.61	0.61	0.65	0.55
	TPR	0.74	0.61	0.70	0.62	0.74	0.61	0.61	1.00
baseline 31	ACC	0.28	0.30	0.34	0.33	0.22	0.24	0.33	0.30
	F1	0.08	0.11	0.11	0.11	0.08	0.09	0.11	0.10
	PPV	0.08	0.11	0.11	0.19	0.09	0.09	0.11	0.10
	TPR	0.11	0.11	0.12	0.12	0.08	0.08	0.14	0.10
baseline 32	ACC	0.30	0.29	0.33	0.36	0.22	0.24	0.33	0.32
	F1	0.12	0.12	0.16	0.14	0.08	0.09	0.13	0.09
	PPV	0.12	0.12	0.17	0.24	0.10	0.09	0.13	0.10
	TPR	0.14	0.13	0.16	0.15	0.08	0.09	0.15	0.10

Table 4. Scoring results for target goalsDiff.

All these consideration that we have done on the results are resumed and showed in the radar plots in Figure 7 for both accuracy and F1 scoring metric, more precisely they illustrate the scores obtained from the test and give us a rapid comprehension of the difference between the classifiers for each target.

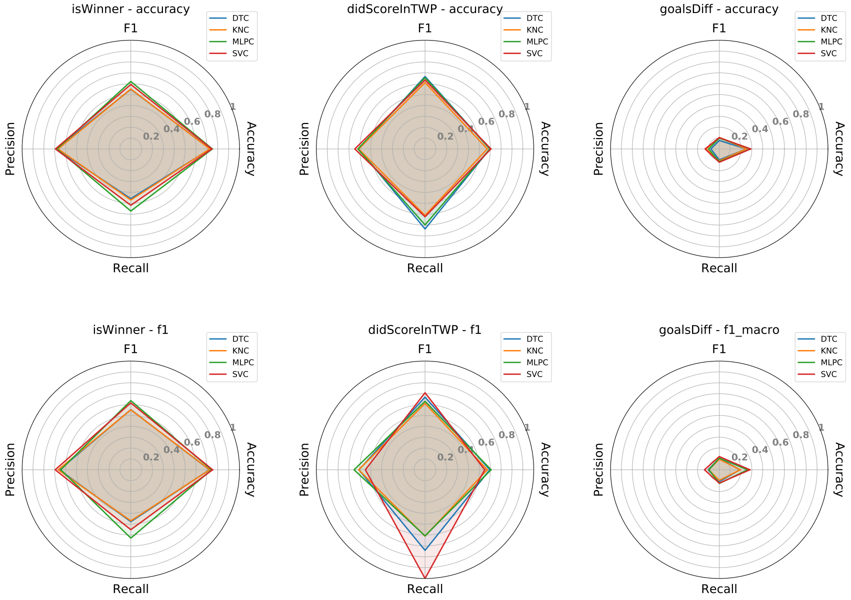

Fig. 7. Radar plots of scores for all the targets with accuracy and F1 scoring.

To resume the observations done on the results of the baseline scores we have plot them in the Figure 8. As said before, for the first target and for both scoring metrics the accuracy of the baseline 11 and 12 is similar, while the results of the other metrics for the baseline 12 are grater than for baseline 11. In the case of second target we clearly see that all the results of F1, precision and recall are zero for the baseline 21. Finally, for the third target, considering the accuracy scoring, all the metrics have the similar gap between the results of baseline 31 and baseline 32.

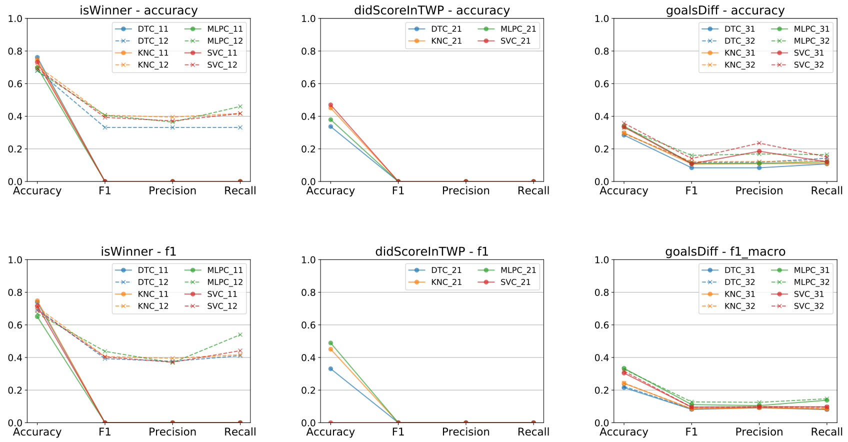

Fig. 8. Plots of baselines scores for all the targets with accuracy and F1 scoring.

3 Explainability

Although the best results were produced by the multi-layer perceptron (MLP), we decided to sacrifice the accuracy for the explainability. As a result, we ended up using the decision tree classifier for which we had reported an accuracy quite close to the accuracy of his well-known rival: MLP.

After the decision was final, we decided to have a moment with our glorious winner and run more experiments to find out its highest potential. After hours of hyper-parameter tuning using grid search algorithm and even manually tuning in some stages, we managed to outperform all of the previous algorithms including the highly fancy ones. Of course this could have happened for the other models given the time and energy. For the purpose of explainability, we wanted our model to be as clear to grasp as possible. For this to be possible, the depth of the tree was chosen to be three. Moreover, it was more desirable to have concrete nodes to avoid overfitting and reduce the number of decision rules. Therefore, we eliminated the nodes in which we had less than 10 percent of the data.

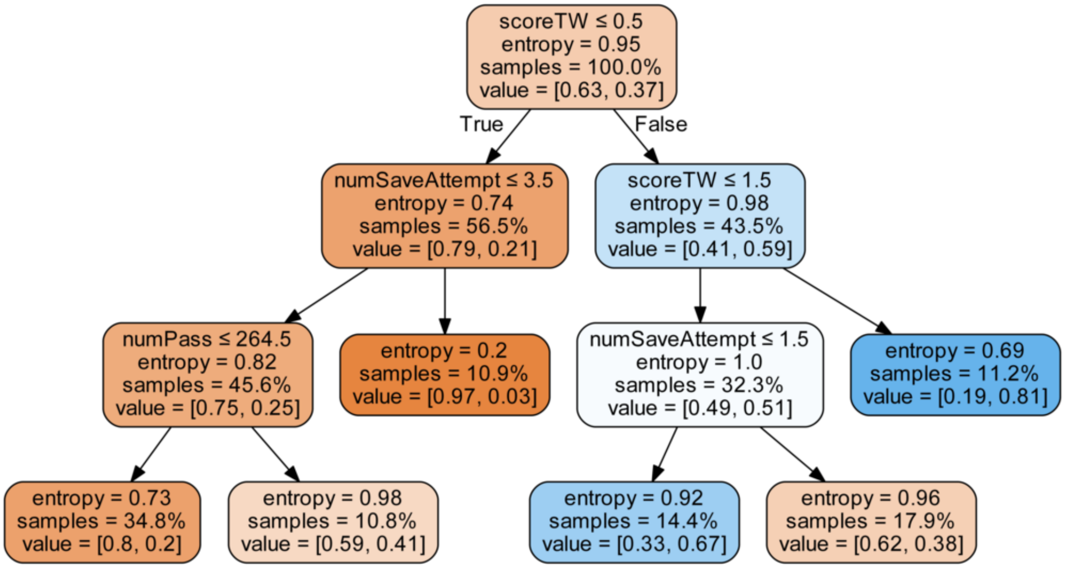

Fig. 9. Decision Tree Classifier trained with the "isWinner" target variable. Blue and Orange are respectfully representatives of the "true" and "false" class.

Looking at 9 gives the user a clear understanding of the logic behind the model's decision making process. This hierarchical process can also be visualized in terms of some specific rules to have in mind. These rules comes directly from the decision tree itself.
Obviously, there are two other target variables for which you can find the best decision trees and the best decision rules on our GitHub repository. Those where decided to be left out of our presentation for the purpose of brevity.

4 Users and Applications

This study can be used in various situations serving various users reaching different goals. It only seems right to mention a handful of these users and applications to induce some ideas in the mind of the curious audience to carry on the future work.

4.1 Coach

The first users of this piece of software could be considered to be the coaches for which the technical evaluation of the game is a task of great importance. A highly accurate soccer prediction tool can assist a coach in different technical decision makings. For instance, in the half time of

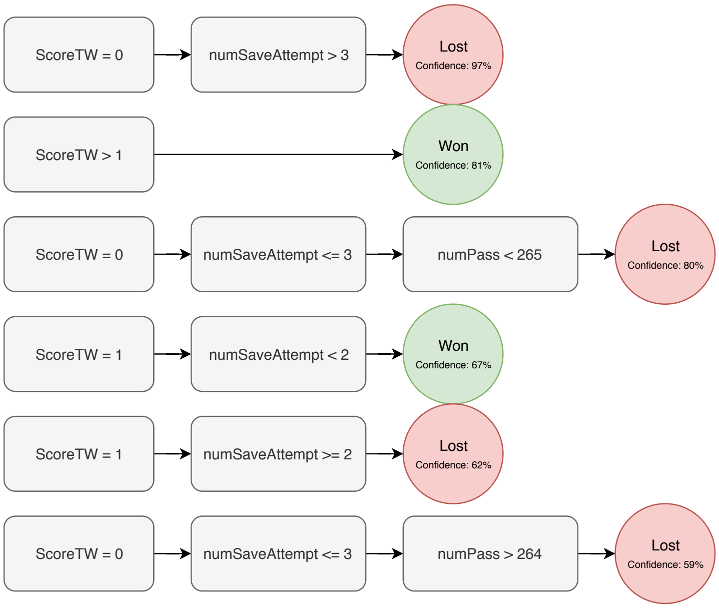

Fig. 10. Decision Rules extracted from the decision tree portrayed in 9 with with the process of finding out the winner gets more and more intuitional.

a losing game the aforementioned coach can take the invaluable players out to save their resources for the following, more promising games. Another application could be the use of the algorithm to understand the outcome of different teams and matches in the same competition to come up with the best strategy for the following or current games.
It goes without saying that coaches usually need the model to be transparent so that there is a valid explanation behind the result being predicted as it is being predicted.

4.2 Non-technical User

Many people all over the world gamble on soccer games. Our application can be particularly interesting to those who want to play safe and rely on the good field of machine learning and statistics to gamble at least more reasonably. It is worth mentioning that the aforesaid user is in no need to have a transparent model. Thus a black box model with higher accuracy could be more desirable to provide here.

4.3 Prosecutor

Fraud detection can be a great deal when it comes to soccer games. However, what makes a fraud claim convincing for the prosecutor?
Looking at the probability of the winning can be a good starting point for the people in charge to start the investigation in case of an unusual

incident where there is a claim stating a case of cheating. Obviously, having transparency can add some extra information to the detectives to work with.

5 Conclusions

Recent years have seen rapid growth of data available on actions taking places during football matches and this technological advancement makes possible to train machine learning models able to improve the performances of a certain football team. For example, a coach may be interested in knowing the most likely outcome of the concurrent match taking place somewhere else (in the same competition) in order to plan the most efficient playing strategy (and possible benches) for his team.

From the classification analysis, and in accord with all the feature that we have implemented or manipulated, we are not able to select one of the four classifiers as the best one. Considering also the correlation matrix that we have produced after the data understanding phase, we imagined that the prediction of these three targets would have been difficult and with useless results.