# Machine Learning methods applied on Soccer Prediction

**Felipe Augusto Pereira Fernandes** FELIPEAPFERNANDES@GMAIL.COM
CEFETMG

## Abstract

Soccer is considered the most popular sport in the world (de Football Association et al., 2007) and therefore it is very common to see people talking about who is going to win a match in family reunions and pub's tables. Looking into a team result history it is possible to establish a performance history along the season. This study presents eight input models for multi layer perceptron artificial neural networks in order to predict the outcome of a soccer match: home victory, draw or away team victory. From basic performance data and inputs that try to capture the influences of the dynamic of a soccer match, this study aims to, at least one proposed model, achieve a mean hit rate greater than 53.25% Using the matches of the 2013-2014 and 2014-2015 English Premier League , seven out of eight models achieved a hit rate greater than 53.25%, with LsM model presenting the higher mean hit rate, 57.5%. These models demonstrate to be an excellent tool to aid soccer match analysis. This study also contributes to encourage new researches to study the soccer match prediction field, which, at the moment, presents a lack of study due to high random variables of a soccer match.

**Keywords**: Soccer match prediction, Pattern Classification, Artificial Neural Networks, Football.

## 1. Introduction

Soccer is the most popular sport in the world. Its popularity can be assigned to its simplicity, because only something that characterize the goal and the ball is needed to its practice. Soccer is also a subject always present on pubs tables, family and friends reunions. Who is going to win the league? Which team is better? are question raised on these talks. This is the reason that nowadays communication vehicles, like TV shows, have now specialists to discuss and make projections about which team is going to win the league or the round's fixture outcome.

Determine the outcome of a match is not a simple task, because soccer is a collective game subject to inumerous variables, such as team pattern, tactic scheme, formation, individuals failures, player's skills, team motivation, referees's decisions *etc*.

Statistics of game, teams and players are used to aid the determine soccer matches's outcome. Nowadays, these statistics are being used on soccer transmission to demonstrate better the scenario of a match, avoiding the subjectivity. Looking into a team, it is possible to establish a performance history along the season. With the help of these statistics numbers such as number of goals per match, percentage of complete passes per game, it is possible to build a numeric chain that is repeatable between matches. As they are repeatable, it could exist a pattern on them.

Artificial neural networks multilayer perceptron are used in solution of pattern recognition and class classification. Its functionality is directly linked with its learning algorithm, that will adjust a neuron's weight, and its architecture.

This study aims to elaborate input models for multilayer perceptron networks with basic performance data, like number of victories, losses, draws, goals scored and goals suffered.

## 2. Previous research

Despite the great popularity, there are few scientific work that address forecast of soccer outcome(Aslan & Inceoglu, 2007). Cheng et al.(Cheng et al., 2003) demonstrate artificial neural networks's good performance when applied to forecast soccer match outcome. In this work, the implemented networks are compared to statistic methods, with higher hit rate.

The figure 1 demonstrates the architecture model proposed by Cheng et al. (Cheng et al., 2003). It is based on a compound, hybrid, system that firstly evaluates teams quality of a match, separating into three classes: home team

*Machine Learning Class – CEFET-MG*, Belo Horizonte, Brasil, 2016.

stronger, denominated Stronger BP on figure 1; teams with same quality- Matchable BP; home team weaker than visiting team - Weaker BP. After separating the data, they are used as input for three Artificial Neural Networks (ANNs) created specific for this data.

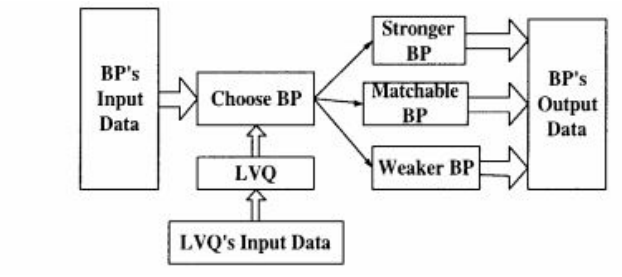


*Figure 1.* Architecture model proposed by Cheng et al. (Cheng et al., 2003)

In order to evaluate the quality of both teams, the difference of mean points and mean scored goals are used. Every match that would be predicted is an input that are put through this selection that defines which ANN that game belongs. The input model is shown on figure 2. $Team(A)$ is the home team and $Team(*)$ is the away team.

$$\begin{bmatrix} Team(A) \\ Team(*) \end{bmatrix} = \begin{bmatrix} win-ratio(A) \\ draw-ratio(A) \\ lose-ratio(A) \\ average-goal(A) \\ average-lose(A) \\ morale(A) \\ hom\,e-away(A) \\ win-ratio(*) \\ draw-ratio(*) \\ lose-ratio(*) \\ average-goal(*) \\ average-lose(*) \\ morale(*) \\ hom\,e-away(*) \end{bmatrix}$$

*Figure 2.* Input model proposed by Cheng et al. (Cheng et al., 2003)

An input, after passing through its designted network, have a predicted outcome. This model achieved a hit rate of 52.29%, higher than statistics methods, such as Elo rating (Elo, 1978).

Aslan and Inceoglu (Aslan & Inceoglu, 2007) made a comparative of forecasting models applied on soccer match. They proposed two models at the end of comparison, basing on the home performance for the home team and away performance for the away team. One model only consider the data from home matches for the home team and away data for the away team. The other model considers home and away performance for both teams. The first network achieved 53.25% as hit rate and 51.29% for the other network. They also concluded that for forecasting networks applied on soccer matches, the input modeling is a critical issue.

## 3. Database

For this study a public database is used. This database is available at kaggle (kaggle, 2016). On this database there are more than 25000 matches, information of more than 10000 players from 11 countries with their lead championship. It also have information of team line up with squad formation(X,Y coordinates) and betting odds for home win, draw and away win.

## 4. Modeling

As football is a team sport, susceptible to various randomness in which physical, technical and psychological influence on the performance of a team, measure the quality of a team, making this quality in a number, it is something highly subjective and complicated.

One can divide a team into three sectors: defense, middle and attack. In general, specialists say that a team with good middle and defense sectors tend to concede few goals, whilst teams with good attack and middle tend to score more goals *etc*. Therefore, quality indicators for team's sector seem to be interesting inputs for modeling this problem. With these quality indicators, a comparison is more reliable, since it is less subjective.

In order to get these quality indicators, the EA Sports game FIFA was used. On this game, the players are rated from 0 to 100 in physical, technical and psychological attributes. So these indicators are calculated by the mean of all attributes of the players that belongs to determined sector. For example, the defense indicator is calculated by the mean of all attributes of all defensive players.

It is known that the location where a team plays, home or away, have an influence on its performance, because playing with the support of the fans or under the pressure of rivals fans have an influence on the psychological. Therefore to forecast a match outcome it is interesting to evaluate the home team performance when playing home and visiting team performance when playing away.

Suspensions, injuries and players transfer between teams is something very common during a season. These changes may affect a team performance on a period of time. In order to capture these variations, it is interesting to use team's performance on the last games as an input for our model.

In order to capture the psychological influence of match's location on a team performance, gather indicators from match database only for matches where the location were the same as the match in evaluation.

In relation to players suspensions, injuries and transfers evaluate team's performance over the season, in a global way, does not seem to be enough. Thinking on these indicators on a graph, analyzing the performance on a local way, on the last five games, seems to be interesting, since these values are different when looking to the global and local performance. When looking into a local manner it seems to be easier to predict the next behavior, since it is considering only the last matches played. The figure 3 demonstrates the goals per game indicator in a global manner, over the season. The figure 4 exemplifies goals per game indicator in a local way, considering only the last 5 matches played. These figures demonstrate the difference on value indicator when changing the manner on how to apply it.


*Figure 3.* Goals scored per game by Liverpool on 2014-2015 season.

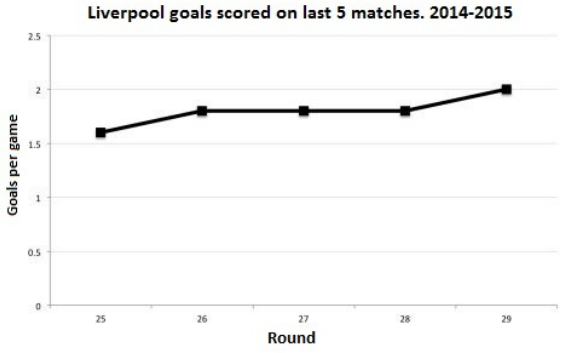


*Figure 4.* Goals scored on the last five games per game by Liverpool on 2014-2015 season.

## 5. Metodology

Firstly, a column selection on matches table is made. Only odds from Bet365 is used, since it is the largest bet company.

Another selection is made, excluding columns of the event incidents. These columns contains a XML that describe that events in a match. The reason of this exclusion is that are few games that these columns are not null. Therefore, as the major matches does not have information on these columns, it is better to exclude them.

A transformation of categorical variables - date, season- into ordinal were made.

In order to predict a match outcome, classification algorithms will be used. So, a target column, Output, is made. This column contains 'H' for home win, 'D' for Draw and 'A' for away win.

This new database is called pure data and it is used on a logistic regression. The features for this first version are : 'country_id', 'league_id', 'season', 'stage', 'date', 'match_api_id', 'home_team_api_id', 'away_team_api_id', 'home_player_X1', 'home_player_X2', 'home_player_X3', 'home_player_X4', 'home_player_X5', 'home_player_X6', 'home_player_X7', 'home_player_X8', 'home_player_X9', 'home_player_X10', 'home_player_X11', 'away_player_X1', 'away_player_X2', 'away_player_X3', 'away_player_X4', 'away_player_X5', 'away_player_X6', 'away_player_X7', 'away_player_X8', 'away_player_X9', 'away_player_X10', 'away_player_X11', 'home_player_Y1', 'home_player_Y2', 'home_player_Y3', 'home_player_Y4', 'home_player_Y5', 'home_player_Y6', 'home_player_Y7', 'home_player_Y8', 'home_player_Y9', 'home_player_Y10', 'home_player_Y11', 'away_player_Y1', 'away_player_Y2', 'away_player_Y3', 'away_player_Y4', 'away_player_Y5', 'away_player_Y6', 'away_player_Y7', 'away_player_Y8', 'away_player_Y9', 'away_player_Y10', 'away_player_Y11', 'B365H','B365D', 'B365A' and 'Output' as the target.

This logistic regression achieved a mean accuracy rate of 52.58%, using a k-fold validation with k=5.

As can be seen, there are two features for each player in a given match. We can turn these 22 features into two features. Just turn them into a formation feature, which contains the number of players in defense, in middle-field and attack.

After this transformation a second logistic regression is made, with the following features: The features for this first version are : 'country_id', 'league_id', 'season', 'stage', 'date', 'match_api_id', 'home_team_api_id', 'away_team_api_id','formation_home', 'formation_away',

'B365H', 'B365D', 'B365A' with 'Output' as target. This logistic regression achieved a mean accuracy rate of 53.15%, using a k-fold validation with k=5.

### 5.1. Feature Engineering

The next step is to perform feature engineering and study its influence on a linear method.

After that, new Machine Learning methods could be applied over the features selected, aiming to improve accuracy rate.

## References


Aslan, Burak Galip and Inceoglu, Mustafa Murat. A comparative study on neural network based soccer result prediction. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*, pp. 545–550. IEEE, 2007.

Cheng, Taoya, Cui, Deguang, Fan, Zhimin, Zhou, Jie, and Lu, Siwei. A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system. In *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, pp. 308–313. IEEE, 2003.

de Football Association, Federation Internationale et al. Fifa big count 2006: 270 million people active in football. *Retrieved February*, 20:2012, 2007.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Elo, Arpad E. *The rating of chessplayers, past and present*. Arco Pub., 1978.

kaggle. Soccer database, 2016. URL https://www.kaggle.com/hugomathien/soccer.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I*. Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.