Prediction learning model for soccer matches outcomes*

Laura Hervert-Escobar†, Timothy I. Matis‡, Neil Hernandez-Gress*

*Instituto Tecnologico y de Estudios Superiores de Monterrey,
{1, 3}laura.hervert@itesm.mx, ngress@itesm.mx
‡Texas Tech University
{2}timothy.matis@ttu.edu

Abstract—Sport is a global business into which passionate fans and smart corporations pour hundreds of billions of dollars. Particularly, football soccer detonates a great movement of money in bets, sponsorships, attendance to parties, sale of t-shirts and accessories, etc. Professional soccer has been in the market for quite some time. The sports management of soccer is awash with data, which has allowed the generation of several metrics associated with the individual and team performance. The aim is to find mechanisms to obtain competitive advantages. In this paper, we propose a procedure for predicting the outcome of a soccer match. The procedure consist of a Bayesian Model based on rank position along with a Machine Learning Model based on historical data of matches. The procedure was tested using a data set containing the results of over 200,000 soccer matches from different soccer leagues around the world and for predicting the outcome of the FIFA world cup 2018. The results showed an improvement in accuracy and rank probability error compared with other methodologies.

Index Terms—Machine Learning; Soccer; Bayesian models; Sport Matches; Prediction

I. INTRODUCTION

The sport is an activity requiring skill or physical prowess and often of a competitive nature. It is performed mainly with recreational objectives and has become an essential part of our lives as it encourages connivance, and when professionally engaged, a way to survive. Within the time, sport has become one of the big businesses in the world and has shown an important economic growth. Thousands of companies have their main source of income in it. Football soccer is the most profitable sport in the world, by quite a significant margin. The most popular sport throughout Europe, Latin America and many parts of Asia, football leagues tend to attract the largest audiences of any professional sport [1]. Combined, the five largest football tournaments in the world (the Premier League, Bundesliga, La Liga, Serie A and Ligue 1) are responsible for approximately £12.29 billion in total revenue on an annual basis. This figure excludes revenue from tournaments such as Campeonato Brasileiro, Major League Soccer and the EFL Championships, all of which add additional revenue to the total earnings of the worlds top football leagues. [2] That is why it has aroused great interest in building predictive and statistical models for it. And now, professional soccer is awash with data due to the time that it has been in the market. Therefore, it is possible the generation of several metrics associated with the individual and team performance. The aim is to find mechanisms to obtain competitive advantages. In this way, Machine learning has become a useful tool to transform the data into actionable insights. Machine Learning is a scientific discipline in the field of Artificial Intelligence that creates systems that learn automatically. Learning in this context means identifying complex patterns in millions of data. The machine that really learns is an algorithm that reviews the data and is able to predict future behavior. It finds the sort of patterns that are often imperceptible to traditional statistical techniques because of their apparently random nature [3]. When the scope of data analysis techniques is complemented by the possibilities of machine learning, it is possible to see much more clearly what really matters in terms of knowledge generation, not only at a quantitative level, but also ensuring a significant qualitative improvement. Then researchers, data scientist, engineers and analysts are able to produce reliable, repeatable decisions and results [3]. With data now accessible about almost anything in soccer, machine learning can be applied in a range. However, it has been used mostly for prediction. This type of models are known as multi-class classification for prediction, an it has three classes; win, loss and draw. According to Gevaria, win and loss are comparatively easy to classify. However, the class of draw is very difficult to predict even in real world scenario. A draw is not a favored outcome for pundits as well as betting enthusiasts [4]. Several studies focus on various aspects of football, from analyzing player development and injury recovery to team psychology and match tactics. This paper is concerned with the challenge of developing a model that is capable of predicting the outcome of future football matches, over multiple leagues and divisions. Main contribution of this research is the cooperative model using history match and bayesian function to obtained a better prediction. The remainder of this paper is organized as follows. Section III gives a summary of previous work on football prediction. A general description of how the problem is addressed is presented in Section IV. Section V describes the procedures for pre-processing data, followed by the description of the

*The authors are grateful to the financial support of CONACYT-SNI program in order to promote quality research.
proposed model. Experiments and results are described in Sections VII and VIII, respectively. Finally, discussion of the results are in Section IX.

II. RELATED WORK

Since soccer is the most popular sport worldwide, and given the amount of data generated everyday, it is not surprising to find abundant amount of research in soccer prediction. Most of related work is focused on developing models for a specific league or particular event such as world cup, with predictions derived using various predictive modelling techniques. These can be divided into statistical models, machine learning and probabilistic graphical models, and rating systems.

Research classified as statistical models are as follows: Crowder [5] proposed a model using refinements of the independent Poisson model from Dixon and Coles. This model considers that each team has attack and defense strategies that evolves over time according to some unobserved bivariate stochastic process. They used the data from 92 teams in the English Football Association League to predict the probabilities of home win, draw and lost. Anderson [6] evaluates the performance of the prediction from experts and non-experts in soccer. The procedure utilized was the application of a survey to a 250 participants with different levels of knowledge in soccer. The survey consist on predicting the outcome of the first round of the World Cup 2002. The results shows that a recognition-based strategy seems to be appropriate to use when forecasting worldwide soccer events. Koning [7] proposed a model based on Poisson parameters that are specific for a match. The procedure combines a simulation and probability models in order to identify the team that is most likely to win a tournament. The results were effective to indicates favorites, and it has the potential to provide useful information about the tournament. Goddard [8] proposed an ordered probit regression model for forecasting English league football results. This model is able to quantify the quality of prediction along with several explanatory variables. Angelini [9] proposed a model based on Poisson autoregression with exogenous covariates and allows to determine the joint probability distribution of all possible match outcomes. The outcome was used to define a suitable betting strategy comparing the probabilities estimated by the PARX models and the odds proposed by the bookmakers. The model was applied to a dataset of the English Premier League.

Research classified as machine learning and probabilistic graphical models are as follows:

Rotshtein [10] proposed a model to analyzed previous matches with fuzzy knowledge base in order to find nonlinear dependency patterns. Then, they used genetic and neural optimization techniques in order to tune the fuzzy rules and achieve a acceptable simulations. Koning [11] used a Bayesian network approach along with a Monte-Carlo method to estimate the quality of soccer teams. The method was applied in the Dutch professional soccer league. The results were used to assess the change over the time in the balance of the competition. Rue [12] analyzed skills of all teams and used a Bayesian dynamic generalized linear model to estimate dependency over time and to predict immediate soccer matches. Constantinou [13] proposed a model designed to predict football match outcomes in one country by observing football matches in multiple other countries. The model is a mixture of two methods dynamic ratings and Hybrid Bayesian Networks.

Finally, rating systems Applications to football match prediction are mainly based on variants of the widely known ELO rating system [14], which was initially developed for assessing the strength of chess players, and include the official FIFA/Coca-Cola World Ranking (FIFA 2017). A rather different rating method, the pi-rating [15], provides relative measures of superiority between football teams solely on the basis of the relative discrepancies in scores between adversaries. Halicioglu [16] analyzed football matches statistically and suggested a method to predict the winner of the Euro 2000 football tournament. The method is based on the ranking of the countries combined with a coefficient of variation computed using the point obtained at the end of the season from the domestic league.

Falter [17]and Forrest [18] proposed an approach focused more on the analysis of soccer matches rather than on prediction. Falter proposed an updating process for the intra-match winning probability while Forrest computes the uncertainty of the outcome. Both approaches are useful to identify the main decisive elements in a soccer league and use them to compute the probability of success.

Among the existing works, the approach of [19] is most similar to ours. Their system consists of two major components: a rule-based reasoner and a Bayesian network component. This approach is a compound one in the sense that two different methods cooperate in predicting the result of a football match. Second, contrary to most previous works on football prediction they use an in-game time-series approach to predict football matches.

III. GENERAL IDEAS

There are several factors that have an impact on the result of a soccer game, such as: the morale and skills of the team and / or player, the training strategy, the equipment, etc. All this makes the prediction of the outcome of a match complex, even for experts in the area. It also raises interesting research questions for a general prediction model that can be applied to any soccer match. For example, how the the rules of each league impact the result of a match? Is it the same to predict a regular league match, a league tournament match or an international tournament? How is it possible to achieve a good prediction by knowing only the results of previous games? In order to answer these questions, a model is proposed to predict the results of soccer games of 52 leagues around the world using around 200,000 results of previous games of regular league games (does not include league tournaments). Details of data set is available at [20]. The same prediction model was used to predict the first phase (group stage) of the FIFA World Cup 2018. In this proposal we use a model supported by two axes, the first is a Bayesian model based
on the ranking of each team and the second is based on the history shared by the teams in dispute. Team ranking is done when the season has been completed. Points for ranking are computed using a combination of FIFA procedure and the total of goals scored and received throughout the season. There is an adjustment period for new teams incorporated in a season. In order to make a fair comparison in the ranking process, the points earned by a veteran team are made up of 20% of their score in the previous seasons and 80% of their score in the current season. The Bayesian model uses the rank position of each team in the match to obtain a probability of success or failure. Subsequently, with the use of random variables generated using a triangular distribution, an adjustment measure is obtained to recalculate the probability to win or lose. When the difference of probabilities in the teams is less than 10%, a draw is declared in the result. The prediction also takes into account the history shared by the teams in dispute. In the analysis of the historical data it was detected that, in some leagues, the teams that face constantly win or lose patterns that are independent of the rank position. A preliminary analysis was applied to a training data set. The results showed a high correlation between the shared history and the result of win, draw or lose a game. Information of Pearson correlation is showed in Figure 1. Additionally, the graphic show the maximum, average and minimum number of repeated matches per league. The graphic is ordered from low repeated matches to high repeated matches. Then, is clear that the average that two teams have been faced from 2005 to 2017 is around 10 times.

In this way, the complete model considers history match patterns along with rank position to determine the probability of winning, losing or drawing a match.

IV. DATA PRE-PROCESSING AND FEATURE ENGINEERING

The information of the data set contains date, season, team, league, home team, away team, and the score of each game during the season. The main objective in pre-processing the data is to set the initial working parameters for the prediction methodology. Then, the metrics to obtain in this procedure are: the rank position of the teams, the start probabilities for the Bayesian function and the shared history between two teams. Preprocessing procedures were easily implemented using R. Equations used during the pre-processing data are as follows. Index \( i \) refers to team, index \( t \) refers to the season of the team playing in the league, finally \( n \) refers to total games played by team \( i \) during season \( t \).

\[
sg_i^t = \sum_{n} (3w_{n,t}^i + d_{n,i}) \tag{1}
\]

Equation (1) describes the computation of the score based on game performance \( sg \). The score computation gives 3 points for each game won \( (w) \) during the season, 1 point for a draw \( (d) \) and zero points for a lost \( (l) \) game. This method is based on the result points from FIFA ranking method. Match status, opposition strength and regional strength are not considered due to the lack of information in the data set.

\[
sb_i^t = \sum_{n} (gf_{n,t}^i - ga_{n,t}^i) \tag{2}
\]

Equation (2) describes the computation of the score based on the number of goals during the season \( sb \). In this way, the score is given by the number of goals in favor \( gf \) minus the number of goals against \( ga \).

\[
gs_i^t = sg_i^t + sb_i^t \tag{3}
\]

\[
\text{score}_t^i = \begin{cases} 
2 (gs_{i-1}^t) + 0.8 (gs_i^t) & t = 1 \\
0.2 (gs_{i-1}^t) + 0.8 (gs_i^t) & t > 1 
\end{cases} \tag{4}
\]

A partial score given in Equation (3) is the sum of Equation (1) and Equation (2). The total score for each season in given in Equation (4). The teams of the league in each season may vary according to promotions or descents derived from their previous performance. As shown in Equation (4), the previous season has a weight of 20% on the total score. The current season has a weight of 80%. In this way, the ranking process takes into account a previous good/bad performance. But it also gives greater importance to the changes that the team makes in the current season. This measure was designed to have a fair comparison between veteran teams playing and rookie teams in the league. In this way, the history of each team will have an influence on their current rankings (whether positive or not) and rookie teams will have a fair comparison that alleviates league change adjustments. The rank of the team \( rank_i^t \) in Equation (5) is given by its position according to the total score. Given a collection of \( M \) teams, the rank of a team \( i \) in season \( t \) is the number of teams that precede it.

\[
rank_i^t = \{ \{ rank_i^t | rank_i^t < rank_j^t \} \} \quad \forall \quad i \neq j, \quad i, j \in M_t \tag{5}
\]

As expected, not all teams are participating in all seasons. Then, missing teams are not considered in the ranking of the current season. Equations (6) and (7) are used to obtained start probabilities to be used in the Bayesian function.

\[
mrank_i^t = 1 - \frac{rank_i^t}{(Max(rank_i^t) + 1)}; \tag{6}
\]

\[
P_{\text{start}}^t = \frac{\sum mrank_i^t}{\sum \sum mrank_i^t}; \tag{7}
\]

Finally, the shared history of the teams is a list that summarizes the number of cases that the same match has been played. The list also contains the probability of win \( pRw_{i-j} \), lose \( pRl_{i-j} \), and draw \( pRd_{i-j} \) a game based on the total matches \( tg \) for a given period. See Equation (8).

\[
pRw_{i-j} = \sum_{n} \frac{w}{tg}_{i-j}; \tag{8}
\]

\[
pRd_{i-j} = \sum_{n} \frac{d}{tg}_{i-j};
\]

\[
pRl_{i-j} = \sum_{n} \frac{l}{tg}_{i-j}.
\]
A procedure for the Bayesian function proposed is given in Figure 2. The procedure starts by computing the prior probability of the two teams in the match (step 1). The team with higher prior probability is labeled as a team, and the team with lower prior probability is subindex as b (step 2). Then, prior probability of the a team is used to generate 1000 random variables using a triangular distribution. $TD[0, 1, prior_a]$ represents a continuous triangular statistical distribution supported over the interval $min = x = max$ and parameterized by three real numbers 0, 1, and $prior_a$ (where $0 < prior_a < 1$) that specify the lower endpoint of its support, the upper endpoint of its support, and the $x$-coordinate of its mode, respectively. In general, the PDF of a triangular distribution is triangular (piecewise linear, concave down, and unimodal) with a single “peak”, though its overall shape (its height, its spread, and the horizontal location of its maximum) is determined by the values of 0, 1, and $prior_a$. Using the random variables, posterior probabilities are computed in step 5. Then, the probability corresponding to mode of posterior is used to compute and adjust measure. The adjust measure is apply to the start probabilities for the next period (step 9). Finally, the probability of win/lose the match in the period $t + 1$, knowing the probabilities of the current period $t$ is given by equations in step 10. This equations correspond to the prior probability based on the adjusted start probability. The procedure for the soccer prediction using Bayesian function and shared history data is given in Figure 3. The probability taken for the prediction model is chosen between two options, shared history or ranking. Either choice allows to update results in the Bayesian function.

The procedure starts by checking the shared history of the match to predict. Based on the total matches, the next step is either use history probability or Bayesian probability. The threshold to decide is set at least 10 games of shared history. This value is taken from the average given in General Ideas section. Then, if the threshold value is greater or equal to 10, the probability lies on previous results. Otherwise, the probability is given by their rank position in the season-league along with the Bayesian function.
VI. EXPERIMENTS

Procedures were implemented on R statistical free license software. In order to prove the value of the methodology the training data set given by [20] was split in two parts for all leagues. First part contains the results from 2000 to 2015. Second part contains data from 2016-2017 and was used as the matches to predict. The metric used in the challenge is the ranked probability score (RPS). The RPS helps to determine the error between the actual observed outcome of a match and the prediction. Description of the metric can be found at [20]. The outcome variables $xW$, $xD$ and $xL$ are in the rank of $[0, 1]$, where the sum is equal to 1. Additionally, the proposed method was tested for prediction of the group stage of the FIFA World Cup 2018.

VII. RESULTS

Figure 4 shows the results of the RPS for the prediction data set of the leagues around the world. As results shows the average RPS = 0.2620. According to this metric, a low value indicates a better prediction. The data set used for prediction is the same data set used by DOLORES algorithm [13]. The procedure tested is different. They proposed a model designed to predict football match outcomes in one country by observing football matches in multiple other countries. The model is a mixture of two methods dynamic ratings and Hybrid Bayesian Networks. Our model analyzes the results per league with a proposal based on Bayesian function and history match. Bars in the figure show the total matches predicted for each league. Results are equal or better than results obtained with DOLORES algorithm. Figure 6 shows the proportion of the method using either Bayesian function and shared history match. As the figure shows, most of predictions were obtained with a cooperative decision. The x-axis moves from leagues using history match to bayesian function. Then circles at the right side used bayesian function, and circles in the left side used history match. Y-axis shows the RPS result, then circles at the bottom side shows a low rank probability error. Finally, the size of the circle indicate accuracy in the prediction, then, a big circle indicates a better accuracy. Results for the FIFA world cup are in Figure 6. The average RPS obtained was 0.2761. The figure also shows an average accuracy of 0.47 in the matches. Groups with the lowest accuracy were Group B with 0.3 and group H with 0.17, the rest of the groups score 0.5 or higher levels of accuracy. The right side shows an average accuracy of 0.69 in the selection of the teams for next stage of the tournament. For example, in Group A, the model indicates that Egypt and Uruguay classified for next stage, when real result classification was for Russia and Uruguay. Some groups were predicted correctly in classification but with a different position. An example is group E, were both Brazil and Switzerland classified for next phase, but it was expected Brazil lead the group.

VIII. CONCLUSIONS

Main motivation of this work was the chance to participate in the call for the soccer challenge as a way to test a basic Bayesian model along with other techniques to predict the outcome of matches in soccer. Despite the lack of knowledge about soccer in general, we were able to first understand the challenge and then developed a prediction model that is easy to implement. From literature reviewed we learned that each league is driven by different motivations that influence the result of a match game. Then, information based only in the result of matches may no accurate allows to recognize useful patterns for prediction. Most of the time inverted in the process of defining the better way of ranking as well as programming the procedures, trying to make them as efficient as possible. The methodology proposed is simply an instance of a more general framework, applied to soccer. It would be interesting to try other sports. In this section, we consider the possibilities for extension. Even though the framework can in principle be adapted to a wide range of sports domains, it cannot be used in domains which have insufficient data. Another approach to explore in the future is a Knowledge-based system. This usually require knowledge of relatively good quality while most machine learning systems need a huge amount of data to get good predictions. It is important to understand that each soccer league behaves according to particular environment. Therefore, a better prediction model should include particular features of the match game, such as the importance of the
game. Availability of more features that can help in solving the issue of predicting draw class would improve the accuracy. Future work in this area includes the development of a model that attempt to predict the score of the match, along with more advance techniques and the use of different metrics for evaluating the quality of the result.

REFERENCES
Fig. 6. Results of FIFA World Cup 2018

<table>
<thead>
<tr>
<th>Round</th>
<th>Average of Prediction</th>
<th>Average of RSP</th>
<th>Accuracy</th>
<th>Group</th>
<th>Prediction</th>
<th>Real</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5625</td>
<td>0.2163</td>
<td>0.6875</td>
<td>A</td>
<td>Egypt</td>
<td>Uruguay</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.5625</td>
<td>0.2488</td>
<td></td>
<td>B</td>
<td>Uruguay</td>
<td>Russia</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.3125</td>
<td>0.3631</td>
<td></td>
<td>C</td>
<td>Argentina</td>
<td>Croatia</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td><strong>Grand Total</strong></td>
<td><strong>0.4792</strong></td>
<td><strong>0.2761</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


