Comparison between a Bayesian Network based soccer predictive model and 538 soccer prediction model

Andrea Cristiano
Master's Degree in Artificial Intelligence, University of Bologna
andrea.cristiano@studio.unibo.it

April 29, 2024

Abstract

This study compares the predictive performance of a Bayesian Network based model developed for soccer match outcome prediction against the established 538 soccer prediction system. Using historical match data sourced from 538, our model is trained on team statistics and match conditions to forecast match outcomes. We evaluate the accuracy of our model in comparison to 538's predictions, highlighting the comparable outcomes of the two models.

Introduction

Domain

In the realm of soccer analytics, one of the most studied and relevant areas consists in the prediction of the outcome of each match. In this project we are going to introduce a Bayesian Network model to predict matches whose structure is inspired by some of the Bayesian Network approaches already present in the literature (such as (Constantinou, Fenton, and Neil 2012) (Owramipur, Eskandarian, and Mozneb 2013)) and by FiveThirtyEight(538) own soccer prediction algorithm, one of the most relevant and well known models in this class (Boice 2018a).

Aim

The aim of this project is to analyse the performances of a Bayesian Network based model to predict the result of soccer matches to see if an approach of this kind is useful and reliable for the task.

Method

To train and test the model we used 538's own database (available at (Boice 2018b)) in order to make the comparison between the two models as reliable as possible. The Bayesian Network model was implemented trough the *pgmpy* library and the inference steps were performed trough Variable Elimination. Since *pgmpy* requires the discretization of the values it operates with, we performed the discretization process over three columns of the dataset, which are going to be the input values of our model. To improve the accuracy of the model, for each one of the columns involved, we tested different values for the discretization number.

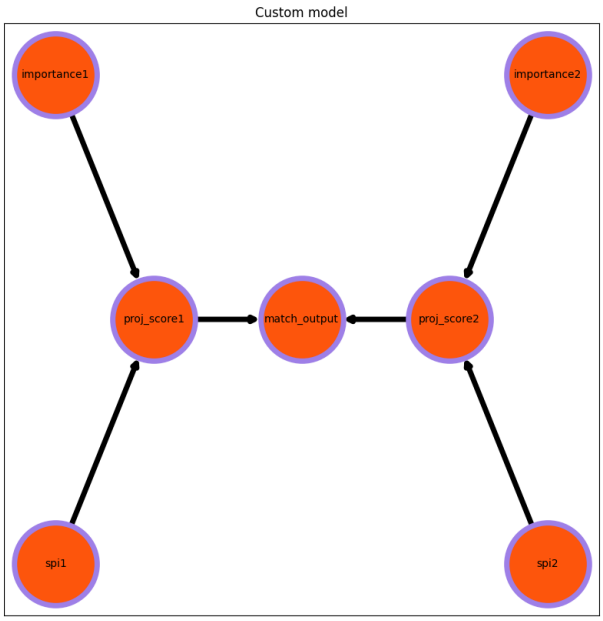

Figure 1: Custom Bayesian Model

Results

We observed that the accuracy of the Bayesian model is comparable to the accuracy of 538's model, for some leagues the accuracy of the custom model is slightly better than 538's, while on others is slightly worse.

Model

The model is composed by 7 nodes in total: 3 for each team and 1 dedicated to the output. The semantic of the nodes is the following:

- Importance: How important is the match for the team in that specific moment of the season, how influential the outcome is on the standings.

proj_score1	proj_score1(1.0)	proj_score1(1.0)	proj_score1(1.0)	proj_score1(1.0)	proj_score1(2.0)	proj_score1(2.0)	proj_score1(2.0)	proj_score1(2.0)	proj_score1(3.0)	proj_score1(3.0)	proj_score1(3.0)	proj_score1(3.0)	proj_score1(4.0)	proj_score1(4.0)	proj_score1(4.0)	proj_score1(4.0)
proj_score2	proj_score2(1.0)	proj_score2(2.0)	proj_score2(3.0)	proj_score2(4.0)	proj_score2(1.0)	proj_score2(2.0)	proj_score2(3.0)	proj_score2(4.0)	proj_score2(1.0)	proj_score2(2.0)	proj_score2(3.0)	proj_score2(4.0)	proj_score2(1.0)	proj_score2(2.0)	proj_score2(3.0)	proj_score2(4.0)
match_output(0)	0.3333333333333333	0.6666666666666666	0.17391304347826086	0.16	0.3125	0.3315580021390374	0.2916666666666667	1.0	0.16981132075471697	0.2101818181818181817	0.0	0.3333333333333333	0.13157894736842105	0.0	0.3333333333333333	0.3333333333333333
match_output(1)	0.3333333333333333	0.16666666666666666	0.11594202898550725	0.04	0.4375	0.3850267379679144	0.20833333333333334	0.0	0.7547169811320755	0.5454545454545454	1.0	0.3333333333333333	0.8157894736842105	0.5	0.3333333333333333	0.3333333333333333
match_output(2)	0.3333333333333333	0.16666666666666666	0.7101449275362319	0.8	0.25	0.28342245989304815	0.5	0.0	0.07547169811320754	0.23636363636363636	0.0	0.3333333333333333	0.05263157894736842	0.5	0.3333333333333333	0.3333333333333333

Figure 2: Conditional Probability Table for Match_output

- Spi: it is a measure of the team overall strength.
- Proj_score: is the projection of the amount of goals we predict each team is going to score given its overall strength and the importance of the match.
- Match Output: it has 3 values describing the probabilities for each possible outcome (0 for draw, 1 for home win, 2 for away win).

Looking at the CPT we can observe that, if the difference between the proj_score of the two teams is significant, then the probabilities in the CPT will reflect that. We also observe that for some combinations the probabilities are distributed evenly.

The structure of the model is based on the basic structure that also 538 uses (Boice 2018a): we use the same input parameters and with those we calculate the projected score, which will, in the end, determine the probability distribution of the output. This allows us to better understand how much the kind of model uses influences the results.

Analysis

Experimental setup

The model has been trained over the seasons 2020-2021, 2021-2022, 2022-2023. The test set is made of the games after the matchday 20. This choice has been made in order to have a sufficient amount of matches to train the model on.

For each match into the test set we took as the model's best guess the proposed outcome of the match whose probability were the highest: we then compared the predicted outcome against the actual outcome. We calculated the accuracy based upon the number of correct guesses the model has made over the total number of match considered.

The inference steps were performed trough Variable Elimination: for each match only the importance and the spi of the two teams were given as an input.

Results

We observed that the model overperforms 538's model in predicting the outcome of Serie A matches and Spanish La Liga, while slightly underperforms 538's in Premier League, League 1, and German Bundesliga. The results satisfy the initial expectations.

Conclusion

In this project we explored the usage of Bayesian Networks as a models to predict soccer matches outcomes. We observed that, using the best combination of parameters, the model returns promising results, averaging an accuracy close to a well-known prediction model.

We can then conclude that Bayesian Networks could be a valuable tool to predict soccer matches outcomes.

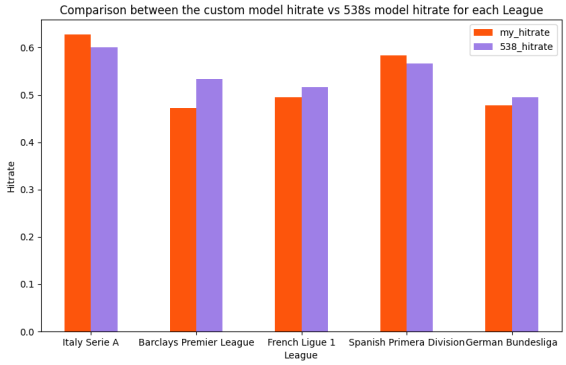

Figure 3: Accuracy comparison between the two models on Europe's top 5 leagues

Links to external resources

- 538 soccer dataset: `https://github.com/fivethirtyeight/data/tree/master/soccer-spi`

References

Boice, J. 2018a. How our club soccer predictions work. `https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/`.

Boice, J. 2018b. How our club soccer predictions work. `https://github.com/fivethirtyeight/data/tree/master/soccer-spi/`.

Constantinou, A.; Fenton, N.; and Neil, M. 2012. pi-football: A bayesian network model for forecasting association football match outcomes. knowledge-based systems, 36, 322-339. *Knowledge-Based Systems* 36:332–339.

Owramipur, F.; Eskandarian, P.; and Mozneb, F. 2013. Football result prediction with bayesian network in spanish league-barcelona team. *International Journal of Computer Theory and Engineering* 812–815.