Human Factors in Software and Systems Engineering, Vol. 61, 2022, 57–66
https://doi.org/10.54941/ahfe1002524

Football Matches Outcomes Prediction Based on Gradient Boosting Algorithms and Football Rating System

Muhammad Nazim Razali[1], Aida Mustapha[2], Salama A. Mostafa[1], and Saraswathy Shamini Gunasekaran[3]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia 86400, Johor, Malaysia
[2]Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia 84500, Johor, Malaysia
[3]College of Computer Science and Information Technology, Universiti Tenaga Nasional, 43000, Selangor, Malaysia

ABSTRACT

Prediction in association football is genuinely a hot topic to discuss as it is among the popular sports that have attracted and gained global interest. The prediction may focus on matches outcomes (win, draw and lose) or the number of goals scored obtained by the home and away teams. This paper proposes football matches outcomes prediction models based on a rating system and gradient boosting algorithms. The testing of the models covers implementing pi-rating and Elo rating as data features generated from limited raw datasets to evaluate match outcomes prediction algorithms such as Gradient Boosting Machine (GBM), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), and CatBoost (CB). The used football dataset has 216,743 instances for learning and 206 instances for testing. The dataset consists of 18 football league seasons between 2001/2002 to 2017/2018 across 35 countries. Subsequently, the prediction results of win, draw, or loss in terms of probability are obtained from the proposed models. The results are compared between several models with different rating systems and different boosting algorithms, as well as past literature that uses a similar dataset. The accuracy and Rank Probability Score (RPS) are set as benchmark criteria. As a result, the pi-rating with CB achieves the lowest RPS, 0.1925, and the highest accuracy of 55.82%.

Keywords: Football prediction, Rating system, Gradient boosting machine, Xgboost, Catboost, LightGBM

INTRODUCTION

Prediction in association football is genuinely a hot topic to discuss as it is among the popular sports that have attracted and gained global interest. The prediction may focus on matches outcomes (win, draw and lose) or the number of goals scored obtained by the home and away teams. According to Constantinou (2019), the prediction models of association football can be divided into statistical models, machine learning and probabilistic graphical models, and rating systems. This division was derived from past

relevant academic studies on association football prediction which focus on leagues or tournament using various of predictive modelling and analysis techniques.

Recently, rating systems tended to be used as part of feature in statistical, machine learning and probabilistic graphical modelling. Constantinou (2019) has modelling hybrid Bayesian network using rating system called pi-rating as well as Hubáček et al. (2019) modelling pi-rating with other relevant features using Gradient boosted trees algorithms. In addition, Robberechts and Davis (2019) apply result-based Elo ratings as part of their features on ordered logit regression and bivariate Poisson model. As the result, the work of Hubáček et al. (2019) and Constantinou (2019) gained 1st and 2nd placed in 2017 Soccer Prediction Challenge (Dubitzky et al. 2019). Meanwhile, Robberechts and Davis (2019) successfully achieved comparable predictive performance with best performing models from the 2017 Soccer Prediction Challenge.

As the time flows, various new algorithms and techniques on predictive modelling developed whether developed from scratches or improved based on previous algorithms which successfully outperforms other traditional techniques or its predecessor in its cluster (Aswad et al. 2022; Nafi et al. 2019). Among of the new techniques are developed and successfully outperforms older techniques in recent years is XGBoost. XGBoost is belong to ensemble method based on gradient booting in machine learning has shown excellent performance in Kaggle's data mining competition by winning 17 out of 29 challenges published in 2015 and even used by every top-10 winning team in KDD Cup 2015 (Chen and Guestrin, 2016). This improvement also influences the predictive modelling in association football. Predictive modelling for football using XGBoost was proposed by Berrar et al. (2019) and gradient boosted trees by Hubáček et al. (2019) successfully gained top-5 in 2017 Soccer Prediction Challenge (Dubitzky et al. 2019). However, the development of ensemble method based on gradient boosting does not stop there when there are more development algorithms such as Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost).

Comprehensive studies on gradient boosting algorithms have been done by Bentéjac et al. (2021) and shows that each gradient boosting algorithms have their own specialty and capabilities in performance analysis. Thus, this paper attempts to analyses and studies football matches outcomes prediction models based on a rating system and gradient boosting algorithms. The remaining paper are organized as follows: Section 2 presents related work on gradient boosting algorithms and rating system that have been used; Section 3 describes the experiments in brief including the dataset, rating system, gradient boosting algorithms and scoring rules used; Section 4 is about results and discussion; Finally, the conclusion in Section 5.

RELATED WORK

The gradient boosting algorithms that has been used in the 2017 Soccer Prediction Challenge (Dubitzky et al. 2019). This challenge was organized for special issue of Machine Learning for Soccer (Constantinou, 2019)

which was participated by several researchers all around the world. Hubáček et al. (2019) and Berrar et al. (2019) has employ gradient boosting algorithms such as XGBoost, GBM and RDN-Boost algorithm to model their football matches outcomes prediction and manage to achieve 1st and 5th places in the challenge. The 2017 Soccer Prediction Challenge is a challenge where the participants to need use machine learning to predict the outcome of 206 future football matches outcomes based on a limited football data describing the match outcomes of 216,743 past football matches (Dubitzky et al. 2019).

Berrar et al. (2019) presented k-Nearest Neighbour (k-NN) and extreme gradient boosted tree (Xgboost) to model their football matches outcome prediction. The k-NN is among the simplest and oldest machine learning algorithms meanwhile the Xgboost can be categorize as the latest and powerful machine learning algorithm since its successfully win many challenges as well as become winning solution of data mining competition such as KDD Cup and Kaggle's Challenge (Chen and Guestrin, 2016). They also presented new ideas on how to integrate domain knowledge of football into modelling process for developing football matches outcomes prediction. These ideas assist them to prepare their limited raw football data into more informative through features engineering. Although their models gained 1st and 5th places in the 2017 Soccer Prediction Challenge in term of accuracy and rank probability score (RPS) as performance metric, they were disqualified since they are the organizers for the challenge.

Besides, Hubáček et al. (2019) which successfully outperform all the competitors' models in the 2017 Soccer Prediction Challenge also proposed two type of gradient boosting algorithms which are Gradient Boosted Trees and Relational Dependency Networks (RDN)-Boost to develop their football prediction models. They run these two algorithms with six difference relevant features selected for developing the learning set include the feature on the historical strength of the team, the current form of the team, pi-rating, PageRank, the match importance and the league specification. As the results, six prediction models were generated namely baseline predictor, relational classification model with and without pi-rating, feature-based classification model, a feature-based classification that only considered pi-rating for prediction and feature based regression model. Thus, the feature-based classification which using all six relevant features successfully achieved the smallest RPS and then this model have been used to participate and won the challenge.

Robberechts and Davis (2019) is difference from Hubáček et al. (2019) and Berrar et al. (2019) since they do not participate the 2017 Soccer Prediction Challenge, however, they used the challenge dataset and problems as benchmark for comparative studies with their proposed football prediction models. They compute and combine result-based Elo ratings and goal-based Offense Defense Model (ODM) ratings and applied ordered logit regression and bivariate Poisson regression to predict football matches outcomes. Thus, their models successfully perform well when compare with best performing models from the 2017 Soccer Prediction Challenge in term of RPS and comparable in term of accuracy.

METHODS AND MATERIALS

The football data, rating system, gradient boosting algorithms and scoring rules will be described in this section. The raw football data extracted as dataset first will be modelled to football team rating system includes Elo rating and Pi-rating and then were learn and test using gradient boosting algorithms include Gradient Boosting Machine (GBM), XGBoost (XGB), LightGBM (LGBM) and CatBoost (CB) for football matches outcomes prediction.

Dataset

The dataset was extracted from Dubitzky et al. (2017) as learning set and Berrar et al. (2017) as testing set which contain 9 features such as season, league, data, home team, away team, home scored, away scored, goal difference and the matches results (Win, Draw, Loss) of 18 seasons of football league between 2001/2002 to 2017/2018 across 35 countries. Overall, learning set consist of 216743 instances and testing set consist of 206 instances. The learning set has increased to 218916 instances after data cleaning including fixing dates, adding labels, completing league data, and removing data duplication (Dubitzky, 2017). Table 1 shows the description of features from the learning and testing set used for the experiment.

Football Rating System

There are three type of prominent rating system have been used in this experiment which are Elo ratings that divided into result-based Elo ratings, goal-based Elo ratings and pi-rating. This rating is computed using learning set data to be utilized for predicting the football matches outcomes using the testing set data.

- **Elo Rating:** Basically, this football team ratings are based on Elo rating system for chess player which then were modified to fit with association football. An Elo rating system represent a single number of football team current strength, where the number of scores obtained by football team can be increase and decrease depend on match result and the ratings of team and its opponents in one specific match. There are two type Elo rating which are result-based Elo ratings and goal-based Elo ratings. The different between result-based Elo ratings and goal-based Elo ratings is that result-based Elo ratings score and calculate the rating based on the result of single match (home win, draw and away win) meanwhile goal-based Elo ratings score and calculate the rating based on the goals difference of single match (the number of goals scores) since a team win by 3-0 are practically more strongly than 2-1 or 1-0 win.
- **Pi-Rating:** Pi-ratings was developed and introduced by Constantinou and Fenton (2013). The pi-ratings computed the football team strength based on home advantage, the current team strength depended on most updated recent results and the win results outcomes are more important than the number of goals difference. The discrepancies rating for team are dependent on rating while played as home or away team, the opponents current rating played as home or away team and the outcome of the match.

Table 1. The description of features from the learning and testing set used for the experiment.

Features	Abbreviation	Description	Datatype
Season	Sea	The season of football league edition	Nominal
League	Lge	The type of football leagues competition	Nominal
Date	Date	The date of the match	Date
Home Team	HT	The team playing at home	Nominal
Away Team	AT	The team playing at away	Nominal
Home Scored	HS	The number of goals scored by home team	Numeric
Away Scored	AS	The number of goals scored by away team	Numeric
Goal Difference	GD	The difference of goals between home team and away team	Numeric
Matches Results	WDL	The outcome of the match in term of win, draw and lose	Nominal

Gradient Boosting Algorithms

There are four types of gradients boosted algorithms have been used in this experiment which are Gradient Boosting Machine (GBM), XGBoost (XGB), LightGBM (LGBM) and CatBoost (CB).

- **Gradient Boosting Machine (GBM):** Gradient Boosting Machine (GBM) is made up of ensemble or combination of weak learner to be used in regression and classification modelling. It typically was run using decision tree as base learner. Though it become one of prominent machine learning techniques, it has many flaws such as it can suffer overfitting, computational expensive and long training time. Howcver, GBM laid foundation for other gradient boosting algorithm development such as XGBoost (XGB), LightGBM (LGBM) and CatBoost (CB).
- **XGBoost (XGB):** Chen and Guestrin (2016) has presented XGBoost (XGB) or also known as Extreme Gradient Boosting Trees and it is the most popular gradient boosting algorithm. It can be employed as solution to many data mining prediction problems and successfully dominate many top data mining competition. XGBoost design as ensemble of decision trees as base classifiers for speed and performance. Literally, XGBoost same in some specification in GBM for principle of gradient boosting, however, more regularized model formalization has been used in XGBoost for controlling or preventing over-fitting which may influence the performance.
- **LightGBM (LGBM):** Light Gradient Boosting Machine or LightGBM (LGBM) was proposed by Ke et al. (2018) to tackle the problem with efficiency and scalability on large size of data that have high dimensional of features. Two novel techniques have been introduced are Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The studies carried out by Ke et al. (2018) shows that LightGBM significantly reduce the computational speed and memory consumption compared to

XGBoost and Stochastic Gradient Boosting (SGB). A comprehensive study done by Bentéjac et al. (2021) on gradient boosting algorithms shows that LightGBM is the fastest compare to Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost (XGB) and CatBoost (CB) however not the most accurate.

- **CatBoost (CB):** Category Boost or known as CatBoost developed by Prokhorenkova et al. (2018). CatBoost is a current newest gradient boosting machine algorithm after XGBoost and LightGBM. One of main advantage of CatBoost is it can handle categorical data features directly without encoding. According to Hancock and Khoshgoftaar (2020), CatBoost is effective and suitable to be employed in various field for many classifications and regression task. The reviews conducted by Bentéjac et al. (2021) on gradient boosting algorithms shows that training speed for CatBoost slower that LightGBM and XGBoost but it succeeds to obtain best results in term of accuracy and AUC accordingly to benchmark dataset.

Scoring Rules

The accuracy and Rank Probability Scores (RPS) has been set as scoring rules in the 2017 Soccer Prediction Challenge (Dubitzky et al. 2019) to access the predictive performance of football matches outcomes prediction models. Thus, accuracy and Rank Probability Scores (RPS) has also been applied in this paper as scoring rules to standardized for comparative analysis of our proposed models against previous competing prediction models such as Constantinou (2019), Hubáček et al. (2019), Robberechts and Davis (2019), and Berrar et al. (2019). The accuracy can be defined as in Equation 1 and RPS can be defined as in Equation 2:

- **Accuracy:** The accuracy can be defined as in Equation 1,

$$Accuracy = \frac{\text{The total prediction results}}{\text{The total observed results}} \tag{1}$$

where the total number of correctly predicted results is divided by the total number of actual observed results.

- **Rank Probability Score (RPS):** Meanwhile the rank probability score (RPS) can be defined as in Equation 2.

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(p_j - e_j\right)^2 \tag{2}$$

where r is the number of potential outcomes, p_j is the forecasted probability of outcome j and pe_j is the actual probability of outcome j.

RESULT AND DISCUSSION

The main objective of this paper is to model prediction for football matches outcomes that solely rely on football rating system as feature with gradient

Table 2. The comparative analysis results of the football matches outcomes prediction models.

Algorithms	Rating System	Accuracy (%)	RPS
GBM	Result-based Elo rating	52.91	0.2001
	Goal-based Elo rating	50.97	0.2055
	Pi-Rating	55.82	0.1938
XGB	Result-based Elo rating	53.39	0.2003
	Goal-based Elo rating	51.94	0.2051
	Pi-Rating	54.85	0.1926
LGBM	Result-based Elo rating	51.94	0.2009
	Goal-based Elo rating	52.42	0.2016
	Pi-Rating	54.85	0.1940
CB	Result-based Elo rating	53.39	0.1997
	Goal-based Elo rating	52.42	0.2024
	Pi-Rating	55.82	0.1925

boosting algorithms. The experiments are conducted using the dataset that incorporates 218,916 instances consisting of 18 seasons of the football league between 2001/2002 to 2017/2018 across 35 countries as a learning set and 206 instances as testing set from 52 football league. The scoring rules were accessed via accuracy and RPS on those matches for measuring the predictive performance. The raw dataset was first computed into rating system (result-based Elo ratings, goal-based Elo ratings and pi-ratings) before were train using gradient boosting algorithms (GBM, XGB, LGBM and CB). The current updated rating system are then be inserted to 206 matches as testing set for prediction.

Table 2 shows the comparative analysis results of the prediction models in terms of accuracy in percentage and RPS accordingly to specific rating system namely result-based Elo ratings, goal-based Elo ratings and pi-rating. Although it seems that there is not much difference of predictive performance between gradient booting algorithms, the results show that prediction models produced by pi-rating system are better than Elo rating system whether based on results or goals in term of accuracy and RPS. It is observed that CatBoost is best performing predictive performance in term of accuracy and RPS whether using Elo rating or pi-rating.

Table 3 shows the results of average accuracy and average RPS for overall comparative analysis for football matches outcomes prediction models based on rating system and gradient boosting algorithms. It is observed that gradient boosting algorithms with pi-ratings achieved highest average performance metrics compared to Elo ratings by obtained 55.34% average accuracy and 0.1932 average RPS.

In addition, the results are compared between several models with different rating systems and different boosting algorithms, as well as past literature that uses similar dataset. The accuracy and Rank Probability Score (RPS) are set as benchmark criteria. The pi-rating with CB has been chosen to compare with the past work since it achieved the best performing model in the studies. The pi-rating with CB achieves the lowest RPS, 0.1925, and the highest

Table 3. The overall comparative analysis results of the football matches outcomes prediction models based on rating system and gradient boosting algorithms in term of average accuracy and average rank probability score.

Rating System	Average Accuracy (%)	Average RPS
Result-based Elo Ratings	52.91	0.2003
Goal-based Elo Ratings	51.94	0.2037
Pi-Ratings	55.34	0.1932

Table 4. The comparative analysis results of the best performing football matches outcomes prediction models.

Algorithms	Accuracy (%)	RPS
Berrar et al. (2019)	51.94	0.2054
Hubáček et al. (2019)	52.43	0.2063
Constantinou (2019)	51.46	0.2083
Berrar et al. (2019)	50.49	0.2149
Robberechts and Davis (2019)	51.46	0.2035
pi-rating with CatBoost	55.82	0.1925

accuracy of 55.82%. However, the performance results comparably are still near to other proposed rating and boosting systems of the past literature.

Table 4 shows the results of comparative analysis of the best performing the football matches outcomes prediction models using gradient boosting algorithms with different team rating against existing models in term of accuracy and rank probability score. The findings of this study may help future researchers develop new football match outcome prediction models that can incorporate several new features and existing features.

This paper's finding gives insight into the possible performance improvement of football prediction models by using other advanced techniques to fuse the data and create more informative features. It is also recommended to expand the limited raw data by using domain knowledge of feature engineering process as well as incorporating more relevant key features data related to human factors such as managerial, teams, and individual factors. This data may provide more information to be accessed, such as availability of players due to international call-ups, tournament, injuries, transfer, or suspension (yellow and red cards), and even the changes of the owner of the football teams and coaches. Besides, subjective information through experts such as fatigue, morale, atmosphere, and motivation of football players as well as fans may influence the football teams' performance and, precisely, football matches results outcomes.

CONCLUSION

This paper present football matches outcomes prediction models based on rating systems and gradient boosting algorithms. The rating systems are result-based Elo ratings, goal-based Elo ratings and pi-ratings. The gradients boosted algorithms are Gradient Boosting Machine (GBM), XGBoost

(XGB), LightGBM (LGBM) and CatBoost (CB). The results show pi-rating system with gradient boosting algorithms are better than Elo rating system whether based on results or goals in term of accuracy and RPS by 55.34% and 0.1932. As compared with past literature, the pi-rating with CatBoost achieves the best performance measures and successfully achieved smallest RPS of 0.1925 and highest accuracy of 55.82%.

ACKNOWLEDGMENT

This paper is supported by the Center of Intelligent and Autonomous Systems (CIAS), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM).

REFERENCES

Aswad, F. M., Kareem, A. N., Khudhur, A. M., Khalaf, B. A., & Mostafa, S. A. (2022). "Tree-based machine learning algorithms in the Internet of Things environment for multivariate flood status prediction", *Journal of Intelligent Systems*, *31*(1), 1–14.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2021). "A comparative analysis of gradient boosting algorithms", Artificial Intelligence Review 54(3), pp. 1937–1967.

Berrar D., Lopes, P., Davis, J. and Dubitzky, W. (2017). "The 2017 Soccer Prediction Challenge", https://osf.io/ftuva/.

Berrar, D., Lopes, P. and Dubitzky, W. (2019). "Incorporating domain knowledge in machine learning for soccer outcome prediction", Machine Learning 108(1), pp. 97–126.

Chen. T. and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system", In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Constantinou, A. (2019). "Dolores: a model that predicts football match outcomes from all over the world", Machine Learning 108(1), pp. 49–75.

Constantinou, A. and Fenton, N. (2013). "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries", Journal of Quantitative Analysis in Sports 9(1), pp. 37–50.

Dubitzky, W., Lopes, P., Davis, J. and Berrar, D. (2017). "The Open International Soccer Database", https://osf.io/kqcye/.

Dubitzky, W., Lopes, P., Davis, J. and Berrar, D. (2019) "The Open International Soccer Database for machine learning", Machine Learning 108(1), pp. 9–28.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). "CatBoost for big data: An interdisciplinary review", Journal of Big Data, 7(1).

Hubáček, O., Šourek, G. and Železný, F. (2019). "Learning to predict soccer results from relational data with gradient boosted trees", Machine Learning 108(1), pp. 29–47.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2018). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017).

Nafi, S. N. M. M., Mustapha, A., Mostafa, S. A., Khaleefah, S. H., & Razali, M. N. (2019, September). "Experimenting two machine learning methods in classifying river water quality", In *International Conference on Applied Computing to Support Industry: Innovation and Technology* (pp. 213–222). Springer, Cham.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2018). "Catboost: Unbiased boosting with categorical features". Advances in Neural Information Processing Systems, vol 31, pp. 6638–6648.

Robberechts, P. and Davis, J. (2019). "Forecasting the FIFA World Cup: Combining result- and goal-based team ability parameters", In Machine Learning and Data Mining for Sports Analytics, edited by Brefeld, U., Davis, J., Van Haaren, J. and Zimmermann, A., Switzerland: Springer, pp. 16–30.