Predicting the Outcome of English Premier League Matches using Machine Learning

Muntaqim Ahmed Raju*, Md. Solaiman Mia†, Md. Abu Sayed‡ and Md. Riaz Uddin§
*Department of Computer Science and Engineering, Dhaka International University, Bangladesh
†Assistant Professor, Department of Computer Science and Engineering, Green University of Bangladesh, Bangladesh
‡§Lecturer, Department of Computer Science and Engineering, Dhaka International University, Bangladesh
Email: *muntaqimahmed@gmail.com, †solaiman@cse.green.edu.bd, ‡saeed.robi@gmail.com, §riazdu54@gmail.com

2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI) |978-1-6654-0489-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/STI50764.2020.9350327

***Abstract*—English Premier League (EPL) is the world's most popular football league. Since this is a prominent league, there has been a variety of preceding endeavors both commercially and scholastically to predict EPL match results. In this paper, machine learning, a promising tool of the fourth industrial revolution (Industry 4.0), has been used to introduce a model for predicting the outcomes of EPL matches both in multi-class (home, draw, and away) and in binary-class (home, and not-home) with the last five seasons football matches. We have employed five machine learning algorithms along with different machine learning techniques ranging from data pre-processing to hyper-parameter optimization which find the best results. In addition, the comparative results demonstrate that, our proposed model gives 70.27% accuracy in multi-class and 77.43% accuracy in binary-class compared to the best known existing models in the literature.**

***Keywords*—Football Prediction, English Premier League, Machine Learning, Data Mining**

I. Introduction

English Premier League (EPL) with a potential television audience of 4.7 billion people is the highest tier of the English football league system and the world's largest sports community. There is a great deal of madness among the people about it and that is the reason of predicting the outcome of EPL matches with a huge phenomenon among football fans. Hence, a lot of football supporters and experts have been giving predictions in different ways about who is going to win before the match begins. Actually, there is a whole industry around it, there are pre-match analysis and post-match analysis by commentators to anticipate who is going to win. Channel like ESPN are committed to try to predict who is going to dominate a game. This is actually very crazy thing that has been going for until the end of time.

Artificial Intelligence (AI) is the brain behind Industry 4.0. Machine Learning, a subset of AI has emerged as a promising tool in intelligent predictive applications for smart manufacturing in Industry 4.0. Thus, in this paper, we have used an intelligent machine learning predictive model which attempts to anticipate who is going to win. The predominant objective of this work is to accurately determine the outcome in multi-class and binary-class of EPL matches. Initially, a survey of the last five seasons of the English Premier League has been conducted. Since then, we have explored a wide variety of soccer blogs, pre-match, and post-match analyses to find out key factors for predicting football match results. We have employed the feature-engineering technique to create the most substantial features. Thereafter, the feature scaling is carried out by using min-max normalization to scale all the features. Uni-variate feature selection based on Chi-Square statistical test is used as a feature selection method for choosing the best viable features set primarily based on their scores for their correlation with the outcome variable. Then to perceive the most promising method of the prediction, we have monitored five different machine learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB) classifier, Decision Tree Classifier (DTC) and AdaBoost Classifier (ABC). Lastly, the hyper-parameter optimization technique is used to achieve the best possible hyper-parameters of every model.

A variety of experiments to forecast football matches have been carried out in the literature. In the exploration [1], the authors discussed the prediction of football matches using tree-based model algorithm such as C5.0, random forest, and extreme gradient boosting and the best accuracy is generated by the random forest algorithm which is 68.55%. However, the primary downside of this analysis is its feature collection. Exploration [1] can not be used to determine football matches prior to the game starts because they used features such as home team shots, away team shots, home team corners, away team corners, etc. which can not be addressed before the match started. Whereas, the accuracy of our study is higher than the study mentioned and is capable of predicting football matches before the game starts. Some studies [2], [3] have been conducted using different algorithms but the predictive accuracy is low, i.e., only 59% and 58.5%, respectively. In another research [4], the output of football matches has some limitations. The algorithm used is the LR that gives only two results, i.e., home or not-home while in a football match there are three possible outcomes home win, away win, or draw.

In this paper, we will discuss prior works before analyzing feature selection, discussing performance of various models, and analyzing our results.

II. Literature Review

Various examines have been done to find the criteria for foreseeing the result of football matches to be more exact.

978-1-6654-0489-1/20/$31.00©20XX IEEE

The following investigations have been led to locate an ideal model for the prediction of football matches.

Alfredo et al. [1] discussed the football match prediction using tree-based model algorithms such as C5.0, random forest, and extreme gradient boosting. The backward wrapper method was used as a feature selection methodology to assist in picking the best feature to improve the accuracy of the model. This study used 10 seasons of EPL football matches history with 15 initial features to predict the match results (home win, away win or draw). The random forest algorithm generated the best accuracy of 68.55% whereas the C5.0 algorithm had the lowest accuracy of 64.87% and the extreme gradient boosting algorithm provided 67.89% accuracy.

Sathe et al. [2] prepared dataset to predict the outcome (home win, away win or draw) of EPL matches by web crawling of team ratings from sofifa and considering the performance of each team at home field and away field. Their final dataset consists of FIFA ratings of each team along with their performances of last 10 seasons. They used three machine learning classification methods, which are Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF). The best accuracy obtained is 59% with SVM method.

Similarly, Baboota et al. [3] worked on the building of a generalized predictive model for predicting the results (home win, away win or draw) of the English Premier League. They used data from 2005 to 2016 spanning of 11 seasons. They divided their dataset into nine seasons of training data from 2005 to 2014, and kept the remaining two seasons from 2014 to 2016 as test data. Using feature engineering and exploratory data analysis, they created a feature set for determining the most important factors for predicting the result of a football match, and consequently created a highly accurate predictive system using machine learning. Their best model using gradient boosting produced accuracy of 58.5%.

Rana et al. [4] described a Logistic Regression model to predict matches outcome (home, not-home) of the English Premier League. They used SVM, XGBoost and Logistic Regression classifiers for primary classification of the data, and then selected the best algorithm out of these three to predict that appropriate label. The application of these classifiers is done on real team data which is gathered from football-data.co.uk for the seasons ranging from 2003-04 to 2018-19. Prediction accuracy of the built model is 65.63%.

III. Goal of the Study

The main goal of this work is to create the most influential features through feature engineering to accurately determine the outcome in multi-class and binary-class of EPL matches. None of the existing works mentioned in Section II worked for both multi-class and binary-class. Since football is a very adaptive game, we have designed our model in such a way that has added very recent data to the model. It will be possible to predict every new season with the help of most influential features.

IV. Research Methodology

In this section, we have presented our proposed examinations of this exploration which employs five different popular machine learning algorithms.

A. Data Collection

The dataset employed in this research originates from DataHub.io, which is a typical dataset to be utilized in football match prediction research. The data used is based on five seasons of EPL matches from the 2014-2015 season to the 2018-2019 season. The total number of data used for this whole investigation is 1870 historical match data.

B. Data Preprocessing

Dataset used in this research needed to be preprocessed since it was composed of several features of each season. Many of these features such as match date, referee name, football team name, and bookmaker odds were practically superfluous. In this process, our essential assignment was to remove the irrelevant attributes or features which had no impact on the model development and keep only the attributes or features we particularly needed. From the retained attributes, feature engineering was done to make the final features that were utilized for model advancement.

1) Feature Engineering: Feature engineering is an important but labor-intensive component of machine learning applications [5]. To use feature engineering, a model's feature vector is expanded by adding new features that compute based on other properties [6]. The final 23 features that have been established with the help of our retained attributes are the mathematical conversion. Some of the features are given below:

- Home team goals scored per game at home: It is a function of home team goals scored at home and home team match played at home. It helps to predict or forecast the number of goals that may be scored by a home team at home.
- Home team goals conceded per game at home: This is based on home team goals conceded at home and home team match played at home. It allows to estimate or determine how many goals a home team would potentially conceive at home.
- Home team win percentage: It is a component of total win and the total match played by the home team. It gives the possibility to win the home team's future match.
- Away team win percentage: This is a measure of the total win and total match played by the away team. It provides the potential of an away team to win future matches.

2) Feature Scaling: Feature scaling is the technology of standardizing individual features over a defined range [7]. For the scaling intent of this study, we have exercised min-max normalization. It is a technique that scales an element or perception into the range of 0 and 1 [8]. The mathematical equation for Min-Max Normalization is,

$$x_{new} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

3) Feature Selection: Feature Selection is the process of selecting a subset of relevant features which contribute most to prediction variable or output [9]. In this study, we have used uni-variate feature selection [10] based on Chi-Square statistical test which picks up the intrinsic properties of the features. Features with the highest Chi-Square statistical test score are illustrated in Fig. 1.

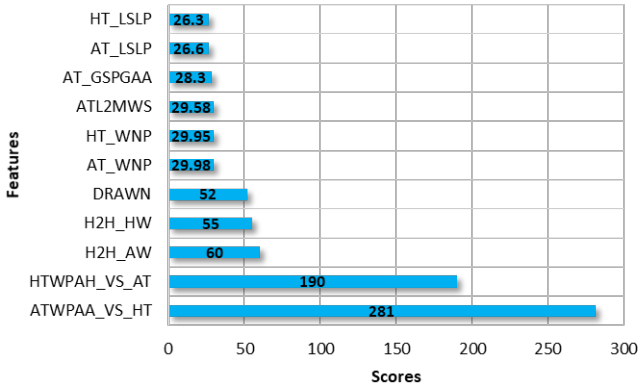

Fig. 1. Features Measured via Uni-Variate Statistics.

Uni-variate feature selection generated the combination of ATWPAA_VS_HT (away-team win percentage at away vs home-team), HTWPAH_VS_AT (home-team win percentage at home vs away-team), H2H_AW (head-to-head away win), H2H_HW (head-to-head home win), DRAWN (draw between home-team and away-team), AT_WNP (away-team win percentage), HT_WNP (home-team win percentage), ATL2MWS (away-team last 2 matches winning streak), AT_GSPGAA (away-team goals scored per game at away), AT_LSLP (away-team last season league points), and HT_LSLP (home-team last season league points) as the set of best features. These features have the potential to have a top influence on the prediction and accuracy of the results.

C. Evaluation Technique

A typical technique for assessment is to split the data into two sub-scales for training and testing. Commonly two-thirds of the data set is used for model building, and one-third is left for testing. However, such single splits often offer consequences that are at danger of sample bias. Many data scientists choose cross-validation in order to decrease bias [11]. In k-fold cross-validation, the dataset is split into k randomly part and each model is trained and tested k times. The cross-validation accuracy is determined by taking the average of the k individual precision measures. Using each fold, the aggregate accuracy CV of the cross-validation is determined A_i.

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{2}$$

D. Models

For the intent of this analysis, we have primarily employed five mainstream supervised machine learning algorithms [12] (SVM, LR, NB classifier, DTC, ABC) to address our classification problem.

1) Support Vector Machine (SVM): The SVM is a kernel-based learning algorithm to address the problem of classification and regression. It produces ideal isolating limits between data sets by resolving a problem of quadratic optimization. The algorithm characterizes the best hyper-plan which divides the number of points with a maximum margin associated with different class names [13]. SVM is a predictive data classification algorithm. So, we have checked out it to take care of our classification issues too. For making a model with SVM, we have taken advantage of 23 features that we have already developed through feature engineering. Thereafter, we have tuned the hyper-parameters using grid search with k-folds cross-validation (we used a k-value of 10) where best hyper-parameters were C = 1, $gamma$ = 0.1, $kernel$ = 'sigmoid' illustrated in Fig. 2.

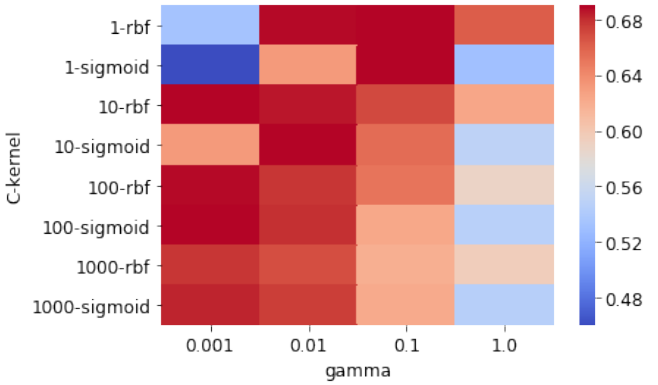

Fig. 2. Grid Search for Support Vector Machine model's hyper-parameters.

SVM predictive model has accomplished a accuracy of 68.99% in multi-class and 76.25% in binary-class where evaluation technique was cross-validation. We then altered the features with features obtained from feature selection which produced 69.15% of accuracy in multi-class and 76.85% of accuracy in binary-class, respectively.

2) Logistic Regression (LR): LR seeks to determine the likelihood of an occurrence on the basis of the independent variables values. It is a statistical method which works with data sets having one or more independent variables that determine the outcome [14]. Our problem is a multi-class classification problem because there are more than two possible outcomes, such as home win, draw and away win. We therefore used multi-nomial LR to construct the prediction model. Grid Search has rendered us the best hyper-parameters for LR model, where C = 0.1, $penalty$ = 'l1' shown in Fig. 3.

LR model has achieved 69.95% of accuracy in multi-class, 77.11% of accuracy in binary-class utilizing all features and 70.27% of accuracy in multi-class, 77.43% of accuracy in

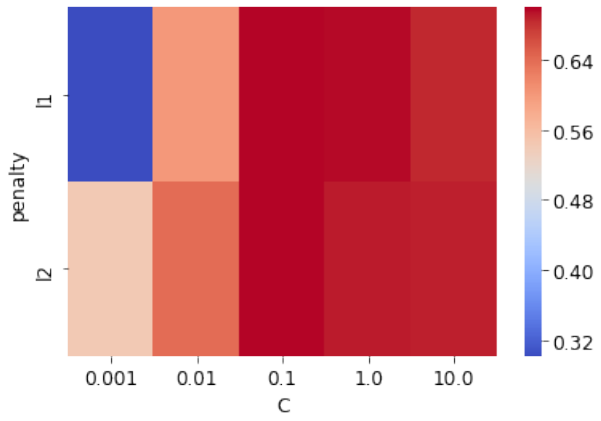

Fig. 3. Grid Search for Logistic Regression model's hyper-parameters.

binary-class using features selected through feature selection strategy, respectively.

3) Naive Bayes (NB): NB algorithm is a machine learning algorithm for classification problems. It is primarily used for text classification, which involves high-dimensional training data sets. It is not only known for its simplicity but also for its effectiveness [15]. We can build models fast and make quick predictions using the NB algorithm. This algorithm learns the probability of an object with certain features belonging to a particular group in class. In short, it is a probabilistic classifier. Probabilistic classifiers are exceptionally equipped for predicting the likelihood distribution. So that, we have set up a probabilistic predictive model for the purpose of our three-class classification problem. For further enhancing the performance of the model, we have optimized the hyper-parameters of the model and the best hyper-parameters were $priors$ = None, $var\_smoothing$ = 0.1 which is illustrated in Fig. 4.

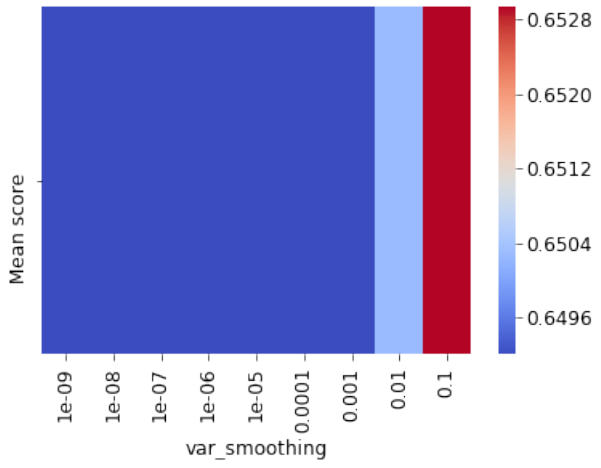

Fig. 4. Grid Search for Naive Bayes model's hyper-parameters.

NB model has obtained 65.95% of accuracy in multi-class, 73.69% of accuracy in binary-class using all features and 67.71% of accuracy in multi-class, 74.92% of accuracy in binary-class using features chosen through the feature selection system, respectively.

4) Decision Tree Classifier (DTC): The DTC algorithm represents a function that takes as input a vector of attribute values and returns a decision single output value. A decision tree reaches its decision by performing a sequence of tests [16]. It can be used to solve both regression and classification problems. Since our problem is also a question of classification, we have built up a predictive model utilizing DTC. We have tuned its hyper-parameters to manipulate the learning process using grid search where best hyper-parameters were $min\_samples\_split$ = 200, $criterion$ = 'gini', $min\_samples\_leaf$ = 1 which is illustrated in Fig. 5.

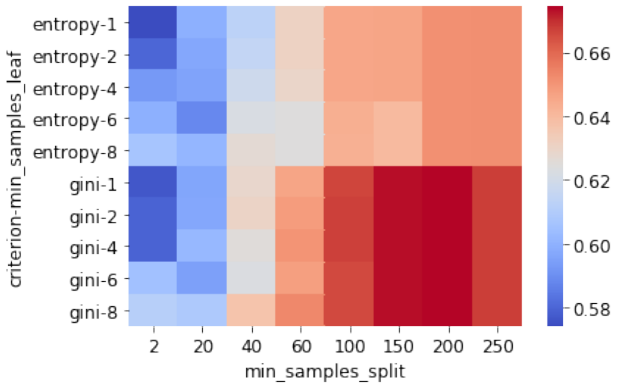

Fig. 5. Grid Search for Decision Tree Classifier model's hyper-parameters.

the DTC model acquired 67.54% of accuracy in multi-class, 74.27% of accuracy in binary-class using all features and 67.76% of accuracy in multi-class, 75.93% of accuracy in binary-class using the features picked by the feature selection system, respectively.

5) AdaBoost Classifier (ABC): ABC consolidates multiple weak learners into a single solid learner. In ABC, the weak learners are single split decision trees, called decision stumps. When ABC stumps its first decision, all results are equally weighted. To rectify the previous error, the incorrectly classified observations bear more weight than the correctly classified observations. ABC is a very powerful boosting algorithm [17]. Not just that, it is commonly used for different problems in machine learning due to its working nature. We have also made a predictive model using it and tuned its hyper-parameters for activating its best level. The most ideal parameters generated by Grid Search were $learning\_rate$ = 0.2, $n\_estimators$ = 80 which is illustrated in Fig. 6.

The predictive model of the ABC has reached towards 68.72% of accuracy in multi-class, 75.02% of accuracy in binary-class using all number of features and 69.15% of accuracy in multi-class, 76.15% of accuracy in binary-class with selected features using the technique of feature selection, respectively.

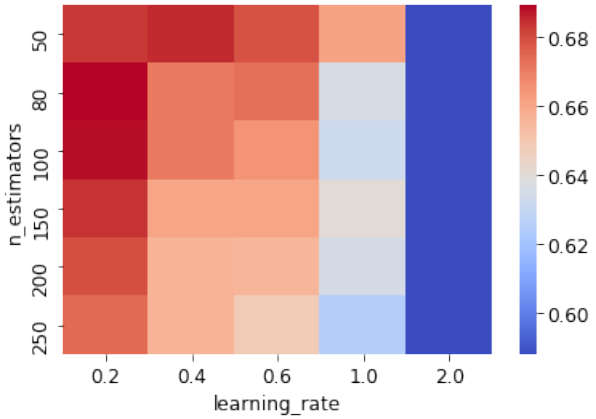

Fig. 6. Grid Search for AdaBoost Classifier model's hyper-parameters.

According to the accuracy rates of each model, the feature selection process has slightly increased the prediction performance.

V. Experimental Results and Analysis

This section exhibits our research findings as well as a comparative analysis with the existing models. Since selected features from feature selection process has slightly increased the performance, we have employed those selected features for the evaluation of each models.

A. Evaluation of Each Model for Multi-Class

The total number of matches was 1870, which consisted of 861 home team wins, 565 away team wins, and 444 draws. The 10-fold cross-validation method with the confusion matrix was executed to measure the efficiency of each classification model. The performance of each model for multi-class is shown from Table I to Table V.

TABLE I
Performance of the SVM Model

Class	Precision	Recall	F1-Score
Away	70%	68%	69%
Draw	43%	68%	52%
Home	82%	70%	76%
Average	65%	68.66%	65.66%

TABLE II
Performance of the LR Model

Class	Precision	Recall	F1-Score
Away	69%	71%	70%
Draw	42%	71%	53%
Home	85%	70%	77%
Average	65.33%	70.66%	66.66%

TABLE III
Performance of the NB Model

Class	Precision	Recall	F1-Score
Away	72%	65%	69%
Draw	46%	59%	52%
Home	76%	73%	74%
Average	64.66%	65.66%	65%

TABLE IV
Performance of the DTC Model

Class	Precision	Recall	F1-Score
Away	66%	66%	66%
Draw	49%	62%	55%
Home	79%	71%	75%
Average	64.66%	66.33%	65.33%

TABLE V
Performance of the ABC Model

Class	Precision	Recall	F1-Score
Away	70%	68%	69%
Draw	42%	68%	52%
Home	83%	70%	76%
Average	65%	68.66%	65.66%

TABLE VI
The Results of the Prediction Process in Multi-Class

Model	Accuracy	Precision	Recall	F1-Score
SVM	69.15%	65%	68.66%	65.66%
LR	70.27%	65.33%	70.66%	66.66%
NB	67.71%	64.66%	65.66%	65%
DTC	67.76%	64.66%	66.33%	65.33%
ABC	69.15%	65%	68.66%	65.66%

According to Table VI, performance values of the Logistic Regression model are a little bit higher than rest of the models. Therefore, we considered LR model as the proposed model of this literature for multi-class classification.

B. Evaluation of Each Model for Binary-Class

The total number of matches was 1870 matches, which consisted of 861 home team wins, and 1009 wins for not-home. To evaluate the efficiency of each classification model, 10-fold cross-validation method was used with the confusion matrix. The performance of each model for binary-class is displayed from Table VII to Table XI.

TABLE VII
Performance of the SVM Model

Class	Precision	Recall	F1-Score
Not-Home	85%	75%	80%
Home	67%	79%	73%
Average	76%	77%	76.50%

TABLE VIII
Performance of the LR Model

Class	Precision	Recall	F1-Score
Not-Home	77%	81%	79%
Home	78%	74%	76%
Average	77.50%	77.50%	77.50%

TABLE IX
Performance of the NB Model

Class	Precision	Recall	F1-Score
Not-Home	74%	78%	76%
Home	76%	71%	74%
Average	75%	74.50%	75%

TABLE X
PERFORMANCE OF THE DTC MODEL

Class	Precision	Recall	F1-Score
Not-Home	77%	76%	76%
Home	71%	73%	72%
Average	74%	74.50%	74%

TABLE XI
PERFORMANCE OF THE ABC MODEL

Class	Precision	Recall	F1-Score
Not-Home	77%	77%	77%
Home	72%	73%	73%
Average	74.50%	75%	75%

TABLE XII
THE RESULTS OF THE PREDICTION PROCESS IN BINARY-CLASS

Model	Accuracy	Precision	Recall	F1-Score
SVM	76.85%	76%	77%	76.5%
LR	77.43%	77.50%	77.50%	77.50%
NB	74.92%	75%	74.50%	75%
DTC	75.93%	74%	74.50%	74%
ABC	76.15%	74.50%	75%	75%

According to Table XII, performance values of the Logistic Regression model are a little bit higher than rest of the models. Therefore, we considered LR model as the proposed model of this literature for binary-class classification.

C. Comparative Results

In this sub-section, a comparative analysis is presented to prove the superiority of the proposed model of EPL match prediction over existing models.

TABLE XIII
COMPARISON OF THE PROPOSED MODEL WITH THE EXISTING MODELS IN MULTI-CLASS

Parameters	Accuracy
Proposed Model in Multi-Class	70.27%
Existing Model [1] in Multi-Class	68.55%
Existing Model [2] in Multi-Class	59%
Existing Model [3] in Multi-Class	58.5%

Table XIII shows the comparison between the proposed model and the existing models [1], [2] and [3] in multi-class where proposed model accuracy is 70.27% and the existing models [1], [2] and [3] have 68.55%, 59%, 58.5% of accuracy, respectively.

TABLE XIV
COMPARISON OF THE PROPOSED MODEL WITH THE EXISTING MODEL IN BINARY-CLASS

Parameters	Accuracy
Proposed Model in Binary-Class	77.43%
Existing Model [4] in Binary-Class	65.63%

Table XIV shows the comparison between proposed model and existing [4] model in binary-class where proposed model accuracy is 77.43% and existing model [4] accuracy is 65.63%, respectively.

VI. CONCLUSION

The model we devised is based on statistical analysis of past football games. We will be able to make fairly accurate predictions. Although the accuracy of this model is pretty good, it is not guaranteed to be always right and there is a lot of scope for future work in this regard. We could bring in sentiment analysis, features such as individual player and team performance metrics, studying the trending hash-tags on twitter on match day, the posts from fans on social media, etc to further enhance the accuracy of the model.

ACKNOWLEDGEMENT

This work was partially supported by the "Research Fund" of Green University of Bangladesh.

REFERENCES

[1] Y. F. Alfredo and S. M. Isa, "Football Match Prediction with Tree Based Model Classification", I. J. Intelligent Systems and Applications, vol. 11, no. 7, pp. 20-28, 2019.

[2] S. Sathe, D. Kasat, N. Kulkarni and R. Satao, "Predictive Analysis of Premier League Using Machine Learning", I. J. Innovative Research in Computer and Communication Engineering, vol. 5, no. 3, pp. 4121-4124, 2017.

[3] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League", I. J. Forecasting, vol. 35, no. 2, pp. 741-755, 2019.

[4] D. Rana and A. Vasudeva, "Premier League Match Result Prediction using Machine Learning", Jaypee University of Information Technology, 2019.

[5] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] A. Coates, A. Y. Ng and H. Lee, "An analysis of single-layer networks in unsupervised feature learning", I. Con. Artificial Intelligence and Statistics, pp. 215–223, 2011.

[7] X. Wan, "Influence of feature scaling on convergence of gradient iterative algorithm", J. Physics: Conf. Series, vol. 1213, no. 3, pp. 1-5, 2019.

[8] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage", IARJSET, vol. 2, no. 3, pp. 20-22, 2015.

[9] J. Tang, S. Alelyani and H. Liu, "Feature Selection for Classification: A Review", Data Classification: Algorithms and Applications, CRC Press, pp. 37-64, 2014.

[10] R. H. Subho, M. R. Chowdhury, D. Chaki and S. Islam,"A Univariate Feature Selection Approach for Finding Key Factors of Restaurant Business", IEEE Region 10 Symposium (TENSYMP), pp. 605-610, 2019.

[11] E. Eryarsoy and D. Delen, "Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods", HICSS, pp. 1107–1115, Hawaii, 2019.

[12] S. Chakravarty, H. Demirhan and F. Baser, "Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting", Applied Soft Computing, vol. 96, pp. 1-17, 2020.

[13] T. Cheng, D. Cui, Z. Fan, J. Zhou and S. Lu, "A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system", Proc. Fifth Int. Conf. Computational Intelligence and Multimedia Applications (ICCIMA), IEEE, 2003.

[14] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", J. Biomedical Informatics, vol. 35, no. 5–6, pp. 352-359, 2002.

[15] D. J. Hand and K. Yu, "Idiots Bayes—not so stupid after all?," Int. Statistical Review, vol. 69, no. 3, pp. 385–398, 2001.

[16] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, "Classification and Regression Trees", Biometrics, vol. 40, no. 3, pp. 874, 1984.

[17] C. Ying, M. Qi-Guang, L. Jia-Chen and G. Lin., "Advance and prospects of AdaBoost algorithm", Acta Automatica Sinica, vol. 39, no. 6, pp. 745–758, 2013.