Collaborative analysis of Football Leagues -English Premier League, La Liga games and Le Championnat

Shirish Sonawane Joseph Cherian Vinod K Mani

*Dublin City University Dublin City University Dublin City University*

[shirish.sonawane2@mail.dcu.ie](mailto:shirish.sonawane2@mail.dcu.ie) [joseph.cherian2@mail.dcu.ie](mailto:joseph.cherian2@mail.dcu.ie) [vinod.mani2@mail.dcu.ie](mailto:vinod.mani2@mail.dcu.ie)

Mustafa Katthawala

*Dublin City University*

[mustafa.katthawala2@mail.dcu.ie](mailto:mustafa.katthawala2@mail.dcu.ie)

Abstract

The advancing field of artificial intelligence has undoubtedly paved the way for a robust predictive system with phenomenal accuracy. With an extraordinary adequacy, machine learning is being exploited essentially everywhere in the best possible way; sport prediction being one among the lot. This paper demonstrates a predictive model for three most celebrated football leagues- English Premier League (EPL), Laliga and Le Championnat. In order for managers and clubs to make the right decision to win tournaments and leagues prediction is very useful [5]. Features that determine the most chances of a team winning the match were investigated and selected, be it home or away. The aspects of feature engineering and exploratory data analysis have been adopted in moulding an accurate predicting model using machine learning. Machine learning models-XG Boost, Support Vector Machine and Logistic regression have been implemented in the prediction to come up with the best among three by evaluation based on their F-score and accuracy.

***Keywords*—** Predictions, Machine Learning, Data Mining, Support Vector Machine, EPL, LaLiga, Le Championnat.

1. INTRODUCTION

Countries world-wide organize football/soccer leagues the duration of which could even be a whole year. More often than not each alliance plays with 20 odd groups chosen from the level divisions and they compete against each other. Matches could either be domestic i.e. home or away and the outcome could be broadly classified as – home win, away win or draw.

Three of the most popular, historically and statistically significant, and widely followed leagues are the English Premier League (EPL), the Spanish La Liga and the French Le Championate. But its fan following does not end within these boundaries and is span across the world with thousands of registered fan clubs having millions of members.

No wonder why the betting agencies and masters of the game spend millions on the analyses of each game given such a huge fan base going gaga over the tournament. It is this analysis of the game statistics that helps not only in better profits of the betting companies, but also focus on improvement of the team formation, player ratings and also for the transfer windows. Most connoisseurs too follow the statistics for discussions happening in sport shows which indeed have immense viewership.

The concept of statistics into sports has been well received irrespective of genre. What followed is the formulation of analysis and predictive capacities alongside which was motivated by the gaining importance of team building. It would help the team managements to formulate strategies once they have an insight of the opponent's tactics which would let them line up with a possible team to tackle. It would make a difference in acquiring potential resources in the transfer market. Alternately the betting companies and reviewers are also spending lots on this to make things interesting and profitable for the fans who would be interested in the forecast beforehand.

By this project, the primary questions we are trying to raise are:

* Can we predict the outcome of EPL, Le Championnat and La Liga matches?
* If we can do the above, how accurate are our results?

2. RELATED WORK

Artificial Neural Network (ANN) is used in critically analysing sport prediction [3], English Premier League here, using ‘SRP-CRISP-DM’ framework for the complex problem of sport result prediction. Though the phases in the model were helpful in shaping our model, we tried to implement algorithms like XG Boost, Support Vector Machine and Logistic regression, they implemented WEKA for modelling. A more relative approach was made by Rahul Baboota andHarleen Kaur, making use of feature engineering and exploratory data analysis using machine learning [4]. Their approach in creating the model was good but lacked enough features and dataset volume. Further, regression-based models were not implemented which we are trying with more volume of data.

Another publication with prediction of soccer match was based on the game day data and current team performance [1] using following 5 classifiers: Linear from stochastic gradient descent, Naive Bayes, Hidden Markov Model, Support Vector Machine (SVM), and Random Forest. They also lacked a good data set and other regression models that we implemented.

A logical regression model had been implemented by the author in predicting results of Barclays’ Premier League results [5]. This work highlighted the importance of features other than home and away status playing vital role in prediction of outcome. We have considered this approach together with other regression algorithms. However, the results obtained were not significant in formulating manager tactics or strategies. A combination of real match data and FIFA game data including player strength, shoot, heading etc. were used by Shin and Gasparyan in their prediction [14]. They claim to have received higher accuracy in prediction by use of game data which also saved much time and resources. They have employed multiple linear and logistic regressions which we inferred for training our model.

A more accurate system implementing Artificial Neural Network (ANN) and logistic regression (LR) techniques with Rapid Miner as a data mining tool claimed to have yielded 85% and 93% prediction accuracy respectively [10]. The aspects of feature selection optimized by weights and their finding on logistic regression which predict only win or loss and no chances of draw were helpful. Though the contents were helpful, we are implementing other regression models as well which are not discussed here. A Bayesian hierarchical model was proposed in another paper, predicting the match outcomes of Serie A matches [2]. What we could gain from this work was selection of feature variable impacting the prediction accuracy. The author used home advantage, team defense and team attack as variables and proved that team with biggest attack won the league. No regression models were used in here that we were focused on.

A research conducted to predict Barclays’ Premier League also attempted in result prediction using logical regression [15]. Again, we could only focus on the significance of variables used here which include the stadium capacity, distance travelled by team, statistics on previous matches, players’ wages, player characteristics etc. The author concluded that most important factors are the previous matches and evaluation of player in attack, defense, midfield and goalkeeper. Another significant work is a multivariate prediction system collecting data from past followed by determining the underlying distribution and significant predictors. Then a link between quality of prediction model in training model and goodness of fit of predictions in holdout sample was drawn. In this multiple linear regression model, variables like travel fatigue, home advantage and ground familiarization were used in prediction which were helpful in building our model.

To interpret parameters affecting crude oil prices and making price estimations, a machine learning model was implemented using XGBoost [16]. It emphasizes that bias variance tradeoff in estimation could be controlled by selecting number of components in model and adding automatic variables in fitting process increases effectiveness in using XGBoost.

3. DATA SET

The data set is from one of the most trusted sources of football data for the leagues (<http://www.football-data.co.uk/>). The statistics of the English premier League, LaLiga League and Le Championate League matches from Season 2000-2001 to 2019-2020 were contained in the dataset in CSV format. Initially each season had its own csv file in each league, so we merged 50 seasons files, which included 20 from EPL, 15 from LaLiga and 15 from Le Championate. Our merged statistical data is having **18700 x 45** (R x C). It is a data set which is comprehensive with different attributes of the game included throughout all these years.

The two kinds of data the dataset is segregated into for analysis are:

**Numerical data** – (FullTime\_HT\_Goal, FullTime\_AT\_Goal, HalfTime\_HT\_Goal, HalfTime\_AT\_Goal, HT\_Shots, AT\_Shots, HT\_Shots\_Target, AT\_Shots\_Target, HT\_Corner, AT\_Corner, HT\_Fouls, AT\_Fouls, HT\_Yellow\_Cards, AT\_Yellow\_Cards, HT\_Red\_Cards, AT\_Red\_Cards')

**Categorical data** – (HomeTeam, AwayTeam, Referee, FullTime\_Res)

Data Dictionary– [**http://football-data.co.uk/notes.txt**](http://football-data.co.uk/notes.txt)contains the full form of the entire dataset abbreviations. It contains the various column descriptions that have been used in the later stages of the paper.

Data collected is mainly includes all the statistics of matches played in Home ground and Away ground. ‘FullTime\_Res’ being the final result of each match, it contains value as ‘H’ (game won by Home team), ‘A’ (game won by Away team) and ‘D’ (No result/Draw).

4. DATA PRE-PROCESSING

*4.1 Handling Missing Values*

While working on final data set, we first tried to figure out with Null values. Our data had total 7 rows with missing values in a total of 18700 rows, i.e., nearly 0.037% null value row in complete data set. Thus, we just discarded the rows with missing values.

*4.2 Dropping Columns:*

External factors like the betting odds shouldn’t affect the predictions. In the datasets the betting data from several organizations have also been included. The data in most of those columns were incomplete. As a result of this the following columns were not included in the prediction:

B365H – Bet 365 Home Win Odds

B365A – Bet 365 Away Win Odds

BSH = Blue Square home win odds

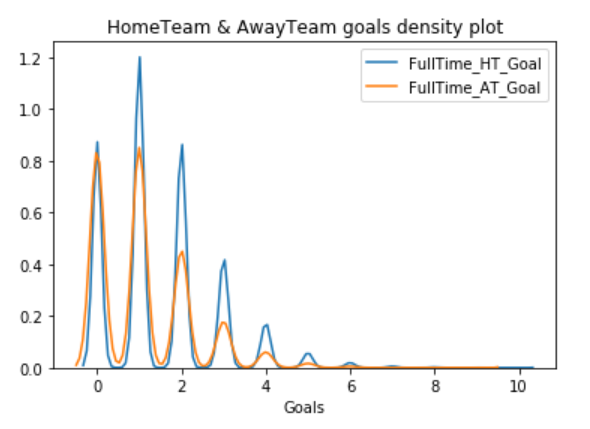
BSD = Blue Square draw odds

B365>2.5 = Bet365 over 2.5 goals

B365<2.5 = Bet365 under 2.5 goals

*4.3 Feature Engineering*

The data set has 45 features which has been recorded for all the seasons from 2000 in the EPL, the Spanish LaLiga and Le Championate. We have included some features in order to improve our understanding and customise our parameters further based on some research into football game statistics.

We initially analysed the two key attributes in our data set ‘Full Time Home Team Goals’ and ‘Full Time Away Team Goals’. Density plot graph is used to get an insight of both the attributes, and we can observe that ‘Full Time Home Goals’ attains peak at 1 and at the same time ‘Full Time Away Goals’ attains two peaks at 0 and 1. Overall continuous distribution view reflects that 80% of both Home and Away goals are less than or equal to 4.

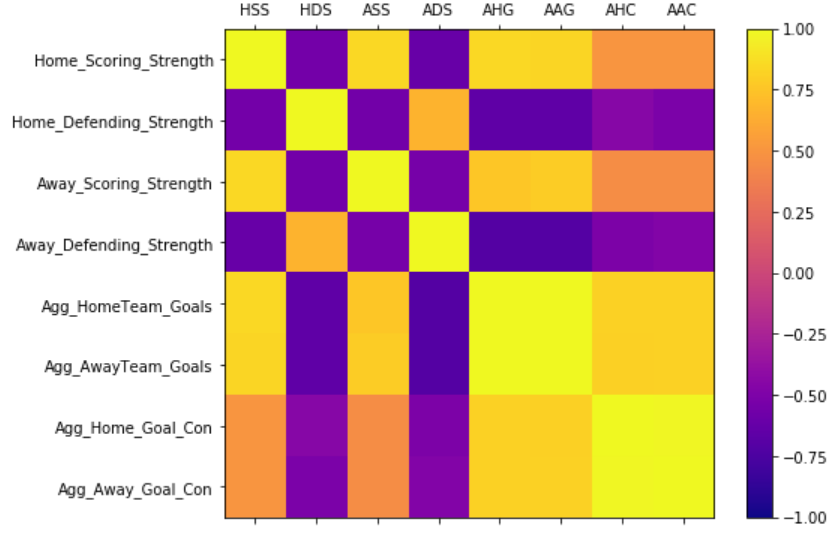
The strength and weakness of individual teams were computed on the basis of overall performance. Below are few computed attributes to understand the correlation between team performance statistics.

* **Matches\_played**- Total Matches played by each team,
* **Agg\_HomeTeam\_Goals**- Total Home goals scored,
* **Agg\_AwayTeam\_Goals**- Total Away goals scored,
* **Home\_Scoring\_Strength**- Ratio of Individual team Home Goals scored average by Total Home Goals average,
* **Away\_Scoring\_Strength**- Ratio of Individual team Away Goals scored average by Total Away Goals average,
* **Agg\_Home\_Goal\_Con**- Aggregate Home Goal Conceded,
* **Agg\_Away\_Goal\_Con**- Aggregate Home Goal Conceded,
* **Home\_Defending\_Strength**- Ratio of Individual team Home Goals conceded average by Total Away Goals conceded average,
* **Away\_Defending\_Strength**- Ratio of Individual team Home Goals conceded average by Total Away Goals conceded average.

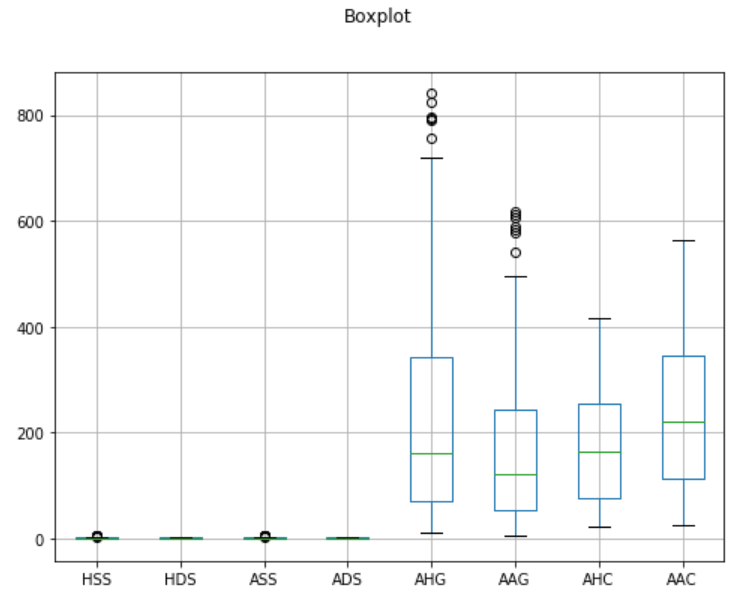
*4.4 Feature Elimination*

A feature table was created by integrating the new features with the existing features.

The feature table contains all numerical data. The FTR column was converted to numerical form (H : +1, A : -1, D : 0) from alphabetical form.

Correlation Matrix for the nine features obtained are shown below. It is perfect for the models to be used as the correlation is very low among most of the features. Correlation Matrix of computed attributes is showing very good correlation between most of the them.

We can observe from Box plot that there are outliers in ‘Aggregate Home Goals’ and ‘Aggregate Away Goals’. The reason being is, we have collaborated data from different leagues and there are few teams from respective leagues who have outperformed in certain games.



5. MODELLING

The following models are included in predictive analysis using the features selected from the feature elimination process.

*5.1. XG Boosting*

The modelling tool for implementing statistical models and analysing data is Python along with ‘Jupyter Notebook’. Keras and respective functional API are used in the easy implementation of all the analysis done above. The two reasons to use XGBoost are:

1. Execution Speed

2. Model Performance

*5.1.2. Model Breakdown*

1. Learning Rate (learning rate = 0.1): XGBoost has a function called as “cv” which performs cross-validation at each boosting iteration and returns the optimum number of trees required. Determination of the optimum number of trees for this learning rate.

2. Max-Depth (max depth = 5): It is the maximum value to which a tree will grow. The overfitting is controlled as higher depth will allow model to learn relations very specific to a sample. Cross Validation is used to tune it.

3. Estimators (n\_estimators = 10): The grid search of the n\_estimators model parameter was performed using scikit-learn. A series of values from 50 to 350 with a step size of 50 (50, 150, 200, 250, 300, 350) were evaluated.

4. Colsample Bytree (colsample\_bytree = 0.3): It denotes the fraction of columns to be randomly samples for each tree. Typical values: 0.5-1.

5. Objective (objective=’reg: linear’): This defines the loss function to be minimized. Multiclass classification using the linear objective, returns predicted class not probabilities. Also set an additional num\_class (number of classes) parameter defining the number of unique classes.

*5.1.3. Feature Statistics*

The feature importance in the model is extracted using the **Trained XGBoost** model. Below is the plot of the 8 features with respective importance.

Results are calculated using F 1 score, in statistical analysis of binary classification, the F 1 score (also F-score or F-measure) is a measure of a test’s accuracy. The precision p and the recall r of the test to compute the score are both considered.

*5.2. Support Vector Machine*

Support Vector Machine (SVM) is a suitable model to be tried on this problem as predicting the outcome of any soccer league is a supervised multi-classification problem. A basic SVM model is built initially on the features selected in the feature engineering process, with dataset split into 70% test and 30 % train and a basic kernel (Kernel = Linear, Gamma = Auto, Regularization Parameter C = 100). An accuracy of 51%, f1-score of 0.43, with 85 % correctly predicted Home win, 2% correctly predicted Draws and 40% correctly predicted Away wins is obtained using this model.

The independent features which are having multicollinearity is considered to have an unstable effect on final estimate. The variance of estimate increases with instability, which means large changes in estimate are caused by a small change in a feature. The few features having high correlation such as Diffpts and DifformPts have been eliminated from the list used to build the final SVM model.

GridSearchCV (Grid Search Cross Validation) function is used for getting the best parameters suited for the SVM model that helps in getting better predictions. The two parts of a GridSearchCV function are Cross Validation and Parameter tuning.The model is trained and tested with different sets of data in **Cross Validation**. Maximizing the accuracy of the model by selecting the values is called **Hyper Parameter tuning**. Kernels: Linear and Radial Basis function, Regularization parameter: 1, 10, 100, 1000 and Gamma values ranging from 0.001 to 0.0001 were the parameters given for **GridSearchCV**. Kernel = RBF, Gamma = 0.001, Regularization parameter C = 1000 were the best parameters selected by **GridSearchCV**.

*4.3. Logistic Regression*

The third model of predictions is a **logistic regression model**. All the 9 features are included in the model as the selected features correlation is very less. As this is a classification problem with multinomial result strategy the model is applied to the target. To fine tune the dimensions available to us we have used ‘**Step-AIC**’ .The final dataset has a training and test set split of 70% and 30% respectively. In order to give the least FTR (False Positive Rate) of 3%, which means a specificity of 70% a threshold is chosen. The final model testing results are as follows.

5. EVALUATION/RESULTS

After running the models mentioned above the following data from the three models was obtained. The confusion Matrix for the models Logistic Regression and SVM are summarized as below.

F – Score and Accuracy Comparison: For the better understanding of the differences between the prediction capacities of all the models we compared the F- Scores of all the model’s vs their accuracies.

Logistic regression Results:

Model used LogisticRegression with training set size 13085. . .

Time taken to train model 0.3796 seconds

Time taken to predict 0.0100 seconds.

1.0 1.0

F1 score and accuracy score for training set: 1.0000 , 1.0000.

Time taken to predict 0.0040 seconds.

F1 score and accuracy score for test set: 1.0000 , 1.0000.

SVC Results:

Model used SVC with training set size 13085. . .

Time taken to train model 6.4275 seconds

Time taken to predict 1.6728 seconds.

0.9989304633019085 0.9990829193733283

F1 score and accuracy score for training set: 0.9989 , 0.9991.

Time taken to predict 0.6528 seconds.

F1 score and accuracy score for test set: 0.9823 , 0.9836.

Linear Regression:

Score is 1

XGB CLassifier report :

Model used XGBClassifier with training set size 13085. . .

Time taken to train model 0.9954 seconds

Time taken to predict 0.0270 seconds.

1.0 1.0

F1 score and accuracy score for training set: 1.0000 , 1.0000.

Time taken to predict 0.0122 seconds.

F1 score and accuracy score for test set: 0.9998 , 0.9998.

6. CONCLUSION AND FUTURE WORK

It is key to ensure that evaluation of each algorithm is done consistently over the data for obtaining a fair comparison. This could be achieved by forcing algorithms to be applied on consistent test harness.

Following six 6 algorithms are compared in the example:

1. Naive Bayes

2. Linear Discriminant Analysis

3. K-Nearest Neighbors

4. Classification and Regression Trees

5. Logistic Regression

6. Support Vector Machines

Each algorithm has been evaluated using a tenfold cross validation procedure and configured with random seed such that same splits to the training data are provided precisely evaluating the algorithms in the same way.

The availability of certain features such as the player information, presence of an important player and player position statistics would definitely yield better prediction results and is not carried out here.[4]

Logistic is the best performing model for the prediction of soccer league results among the 3 models implemented.

7. REFERENCES

[1] Ulmer, B., Fernandez, M. and Peterson, M., 2013. *Predicting soccer match results in the English Premier League* (Doctoral dissertation, Ph. D. thesis, Doctoral dissertation, Ph. D. dissertation, Stanford).

[2] Razali, N., Mustapha, A., Yatim, F.A. and Ab Aziz, R., 2017, August. Predicting football matches results using Bayesian networks for english premier league (EPL). In *Iop conference series: Materials science and engineering* (Vol. 226, No. 1, p. 012099). IOP Publishing.

[3] Bunker, R.P. and Thabtah, F., 2019. A machine learning framework for sport result prediction. *Applied computing and informatics*, *15*(1), pp.27-33.

[4] Baboota, R. and Kaur, H., 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), pp.741-755.

[5] Prasetio, D., 2016, August. Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)* (pp. 1-5). IEEE.

[6] Harrop, K. and Nevill, A., 2014. Performance indicators that predict success in an English professional League One soccer team. *International Journal of Performance Analysis in Sport*, *14*(3), pp.907-920.

[7] Owramipur, F., Eskandarian, P. and Mozneb, F.S., 2013. Football result prediction with Bayesian network in Spanish League-Barcelona team. *International Journal of Computer Theory and Engineering*, *5*(5), p.812.

[8] Timmaraju, A.S., Palnitkar, A. and Khanna, V., 2013. Game ON! Predicting English Premier League Match Outcomes.

[9] Koopman, S.J. and Lit, R., 2015. A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), pp.167-186.

[10] Igiri, C.P. and Nwachukwu, E.O., 2014. An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, *4*(12), pp.12-20.

[11] Louzada, F., Suzuki, A.K. and Salasar, L.E., 2014. Predicting match outcomes in the English Premier League: Which will be the final rank. *Journal of Data Science*, *12*(2), pp.235-254.

[12] Boldrin, B., 2017. Predicting the result of English Premier League soccer games with the use of Poisson models. *Master's thesis, Stetson University, DeLand, FL*.

[13] Kyriakides, G., Talattinis, K. and Stephanides, G., 2017. A Hybrid Approach to Predicting Sports Results and an AccuRATE Rating System. *International Journal of Applied and Computational Mathematics*, *3*(1), pp.239-254.

[14] Shin, J., & Gasparyan, R. (2014). A Novel Way to Soccer Match Prediction. Stanford University:Department of Computer Science.

[15] Snyder, Jeffrey A.L. (2013). What Actually Wins Soccer Matches: Prediction of the 2011-2012 Premier League for Fun and Profit. *Thesis, University of Washington, WA: Department of Computer Science*.

*[16]* Mesut Gumus and Mustafa S. Kiran, Crude oil price forecasting using XGBoost, *2017 International Conference on Computer Science and Engineering (UBMK)*

**8. GITHUB LINK-**

https://github.com/shirishbmc/DataAnalytics/tree/master