Football Prognosis using Machine Learning Algorithm XGBoost

Wasim Gourh[1], Keshav Poojary[1], Mallika Vengarai[1], Nida Parkar[2]

[1]Department of Computer Engineering, Student,University of Mumbai, Atharva College of Engineering, Malad, Mumbai, India
[2]Department of Computer Engineering, Assistant Professor,University of Mumbai, Atharva College of Engineering, Malad, Mumbai, India

Publication Info

Article history:
Received : 14 February 2020
Accepted : 20 May 2020

Keywords:
EPL,FURIA, GBDT, GBM, IDLE, RPS,SVM

****Corresponding author:***
Wasim Gourh
e-mail: *wesgourha@gmail.com*

Abstract

Sport is among the most popular activities of all time. About half of the populace were interested in different activities, football or soccer, as it is popularly called. Football is now an activity with massive venture resources and sales in milliards yearly, not just in terms of sports. In the last year, the Premier League itself has given rise to more than 1 billion dollars. Because betting in most European countries is now allowed, citizens continue to participate week after week. Betting companies have their predictors that are the foundation for betting scores. If people seek to conquer these challenges, they also take a tremendous deal away from them. Even though there may be three outcomes: win, loss, or draw results in a football match, it can be difficult to predict these results. The goal of our paper is therefore to create a model that tries to overcome these odds by utilizing limited data.

1. INTRODUCTION

Sport is a very common and ubiquitous way to spend free time, to be safe and to earn money. Management of sports teams spend huge sums of money on upgrading their athletes and technical equipment and facilities. Collecting participant and match details helps researchers to analyze previous team performances as a unit and deduce effects on team performance from various influences. Well-known is the area of predictive analysis, which is closely associated with athletics. For academics and observers, sports forecasts have always been difficult and football (soccer) is deemed such a phenomenon. There are two areas in general, where forecasts can be made. Gambling companies brokers use statistics to define betting odds [1] for the various teams, players and matches to earn money for their businesses. There are people, on the other hand, who are trying to beat these odds and earn money for successful combinations of single match tips. A match will have three potential outcomes: home win, tie and victory away. Predicting results is a very difficult and relatively straightforward task because of its success and the small number of possible outcomes of the games. However, forecasting the outcome is very complicated, because the way the team performs on a given day relies on many things, such as the current form, the last team meetings, the rivalries, the attack and defense capabilities, the playmaking abilities of the main player, as well as the psychological impact of the fans in the stands [2]. Our research is also focused on forecasting these football match results to create a robust model capable of containing information that could be used to improve the football team's composition.

2. RELATED WORK

J. Hucaljuk and A. Rakipovid have developed a software system that will be able to anticipate the result of the Champions League with about 60 percent precision. To address this, they have first opted for feature selection which is the most important step for predictions. They have selected the following features: the current form of teams shown based on results obtained in the last six months, the result of the previous game-playing squad meeting, the latest ranking place, the number of disabled players on the first squad, the total number of goals earned and earned per season. They performed a series of tests to find the optimal mixture of sets [2].

N. Tax and Y. Jousts. have created two different models, one with the data which are available to the public and other which are to be insider data like injuries to the players, which might be available to the bookies model. Because of the bookmaker's long experience in the task and their commercial interest in the task, betting odds data will be used as a baseline for evaluation of the to be created public data model [3]. The highest accuracy for the public data model was seen when the Naive Bayes or classifier was used in combination with a Components Analysis (with 3 or 7 Components), which achieved an accuracy of 54.702%. The highest accuracy measured using betting odds features was 55.297% when the FURIA classifier was used with 10 folds as parameter.

A generalized statistical software for forecasting the outcome of the English Premier League has been demonstrated by Baboota, Rahul & Kaur, Harleen. (2018). Using software development and exploratory data processing, they have created a software collection to evaluate the most significant variables for forecasting the outcome of a football match and, as a result, create a highly accurate predictive algorithm using machine learning[4]. Their best gradient-boosting model achieved a result of 0.2156 on the graded likelihood score (RPS) metric for Game Weeks 6 to 38 for the EPL aggregated over two seasons (2014–2015 and 2015–2016), while the betting organizations we find (Bet365 and Pinnacle Sports) achieved an RPS rating of 0.2012 for the same period.

Logistic regressions were carried out to determine the correlation between the precision of the forecast scores and the competence of the participants (expert, beginner, lay person), age and gender balance of Khazaal, Y., Chatton, A., Billieux, J. (2012) [5]. 2.3. The variables measured did not affect the precision of the performance prognosis (R2 ranged from 1% to 6%). As a result, experience, age and class did not seem to have an impact on the accuracy of the football game prediction.

3. PROPOSED METHODOLOGY

Football is a sport in which the match length is set, with the team either losing, winning or drawing. In previous models, algorithms have been created to determine the capacity of each team with the help of which the results have been achieved, but today the objectives of the teams are less so as the teams move toward the defensive approach.[5] A factor can also determine whether the team's star player is in the lineup, i.e. the effect of the player on the match. In this paper, together with team results, we seek to include the position of each member and how that player may affect the overall performance of the team and change the result.

3.1. Dataset

Essentially, in the simplest description, when playing football, 11 people on each side play for 90 minutes without advance study of football games. Every player as a human being is special and can be in an acceptable state of mind or not in a very quiet condition. In the sports world, this psychological factor is truly important. So, we have been trying to find a way to measure and calculate this element anyway. These attributes could be the right way to quantify individual players in the whole team and also concerning previous match results[4]. We have collected data of the highest quality and accuracy from multiple sources and this data type comprises the final dataset:

- *Player Statistics:* The most significant stats of the players are the total ranking, ability, and hexagon of abilities that include speed management, shooting, physical strength, running, defensive skills and dribbling skills. The input matrix includes these attributes. Such qualities are measured by scouting and soccer skilled experts who are paid by EA Sports, whose job is to keep all the details up to date according to the real form of the match [2].
- *Match history:* Fulltime and halftime records from 25 seasons back to 1993/94 with up to 22 European league groups. Certain match figures for the major

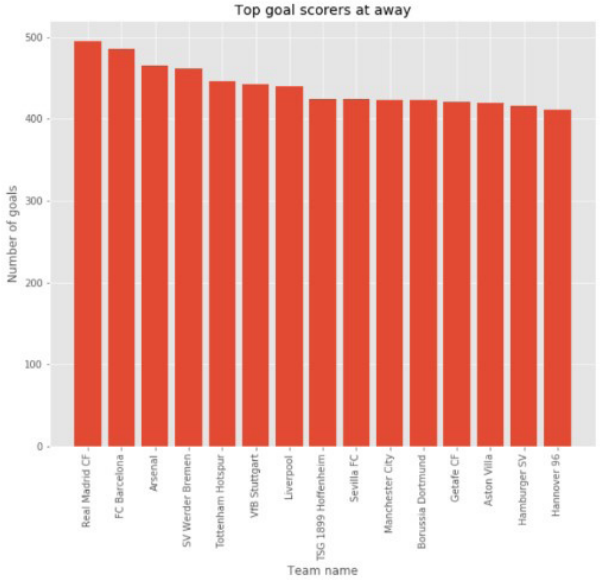

Figure 2: Goal scorers at away

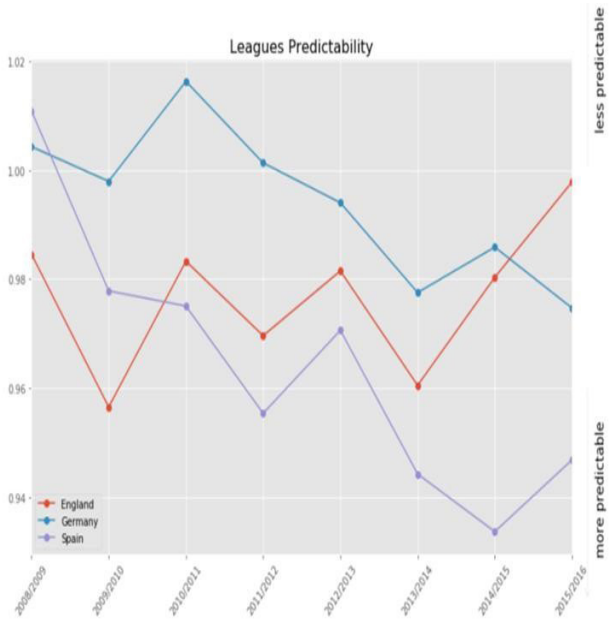

Figure 3: League Predictability

2010-2019 S-JPSET : Vol. 12, Issue Supplementary-1, ISSN : 2229-7111 (Print) and ISSN : 2454-5767 (Online)

European football leagues are also accessible from 2000-01 for goal scoring (includes Corners, foul points, offsides, bookings, red cards, and referees), including British Premiership, Lower Divisions, the Scottish Premiership, German

Bundesliga, Spanish la Liga, Italian Series A and French Championship.

4. SYSTEM DESIGN

4.1. Feature Selection

The end outcome of the football match is decided by a variety of factors, like the nature of the opponents, the home court benefit, the overall team output, the individual quality of players. The data sets proposed and their relationships between them from the previous section are ideal for different combinations of input vectors. Feature selection will be done using sequential forward selection, and best first selection.

4.2. Learning Algorithms

- *Logistic regression:* As in other methods of regression analysis, the logistic regression uses one or more continuous or categorical predictor variables. Nevertheless, unlike traditional linear regression, logistic regression is used to forecast dependent variables, which have inclusion rather than ongoing effects in one specific number of categories. Because of this disparity, linear regression principles have been broken. The residuals cannot usually be dispersed. Therefore, linear regression may establish insensitive projections for a binary variable [7], [8].

Figure 4: Training set 1

- *SVM:* In the light of multiple instances, each classified as belonging to one category or another, the SVM training algorithm constructs a model that provides new instances in one category or another, and this makes it a subtle binary classification that is unlikely [9]. This is the specifically defined differential classifier of the separating hyperplane. In other words, the algorithm provides an optimal hyperplane that categorizes new instances, given the labeled training data (supervised learning).This hyperplane is, in two dimensions, a line that separates the plane into two sections and lies on each side of each segment.
- *Naive Bayes:* Naive Bayes is a basic technique for creating classifiers: models that assign class labels to problem cases, defined as vectors of feature values, where class labels are drawn from a finite set. There is not a single algorithm for the training of these classifiers, but a family of algorithms based on a general principle: all Naive Bayes classifiers presume that the value of a particular feature is independent of the value of every other feature given the class variable [10]. For example, if the fruit is red, round and around 10 cm in diameter, it may be called an apple[2].
- *XGBoost:* Under the gradient boosting framework, it implements master learning algorithms. XGBoost offers a tree boosting parallel (also known as GBDT, GBM), which quickly and accurately resolves many data science problems. The same code is used in large distributed environments (Hadoop, SGE, MPI) and over billions of examples can solve problems. XGBoost has produced better accuracy results compared to other models used in this dataset. The following steps are involved in gradient boosting:
- F0(x) – with which we initialize the boosting algorithm – is to be defined:

$$F(x) = argmin_{\gamma} \sum_{i=1}^{n} L\,(Y_i, \gamma) \quad (1)$$

- The gradient of the loss function is computed iteratively:

$$r_{im} = -\alpha[\partial(L_{\partial F}(v^{i}_{(\ xi\ )}{}^{,F}{}^{(\ xi))}]_{F}(x) = F_{m-1}(x), \quad (2)$$

where α is learning rate

- Each hm(x) is fit on the gradient obtained at each step
- The multiplicative factor γm for each terminal node is derived and the boosted model Fm(x) is defined:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3)$$

5. IMPLEMENTATION

Python programming language has been used to implement this model. The IDLEs Jupyter notebook, as well as PyCharm, have been used for this purpose. Kaggle is a platform for predictive modeling and analytics contests in which businesses and academics post data and statistics

```
XGBoost  :  0.85065789473684222  :  0.85065789473684222
```

In [54]:
```
xt = [[1.05621805792163,0.938271604938271,1.03703703703703,1.15243270868824]]
xt = np.array(xt)
#type(model)
print(model_fit)
model_fit.predict(X_train)
#feature_table
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
              nthread=None, objective='multi:softprob', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

Out[54]:
```
array([ 1,  1, -1, ...,  1,  1,  1], dtype=int64)
```

In [87]:
```
model_fit = XGBClassifier().fit(X_train,y_train)
home_team_ip = input()
away_team_ip = input()
data_home = table_data.loc[table_data["HomeTeam"] == home_team_ip]
data_away = table_data.loc[table_data["HomeTeam"] == away_team_ip]


xt1 = data_home[["HAS","HDS"]].values
xt2 = data_away[["AAS","ADS"]].values

xt = np.concatenate(np.concatenate((xt1,xt2)))
xt = [xt]

#Count elements in xt
count = 0
for listElem in xt:
    count += len(listElem)
count

if count != 4:
    print("Invalid Team Name")

#Prediction

result_predict = model_fit.predict(xt)                #Numpy.ndarray
if result_predict[0] == -1:
    print(away_team_ip + " Wins")
elif result_predict[0] == 1:
    print(home_team_ip + " Wins")
else:
    print("Draw")
```

```
Man City
Man United
Man United Wins
```

In [77]:
```
clf = [MultinomialNB(alpha=1), SVC(kernel = 'linear', C=1.5, probability=True), LogisticRegression(), XGBClassifier()]

labels = [ 'Naive Bayes', 'SVM', 'Log regres', 'XGBoost']

mean_scores = []
mean_scores_2 = []
cms = []

for i in range(0,4):

    clf[i].fit(X_train,y_train)
    clf[i].fit(X_train_2,y_train)

    scores = cross_val_score(clf[i], X_train, y_train, cv=10)
    scores_2 = cross_val_score(clf[i], X_train_2, y_train, cv=10)
    print (labels[i]," : ", scores.mean(), " : ", scores_2.mean())
    #print (labels[i]," : ", scores.mean())

    mean_scores.append(scores.mean())
    mean_scores_2.append(scores_2.mean())
```

```
Naive Bayes  :  0.48421049250120707  :  0.5644751778362285
SVM  :  0.5396360800776296  :  0.6120092569733144
Log regres  :  0.5544398257580837  :  0.5914475668592603
XGBoost  :  0.7207155037511408  :  0.8478625940347942
```


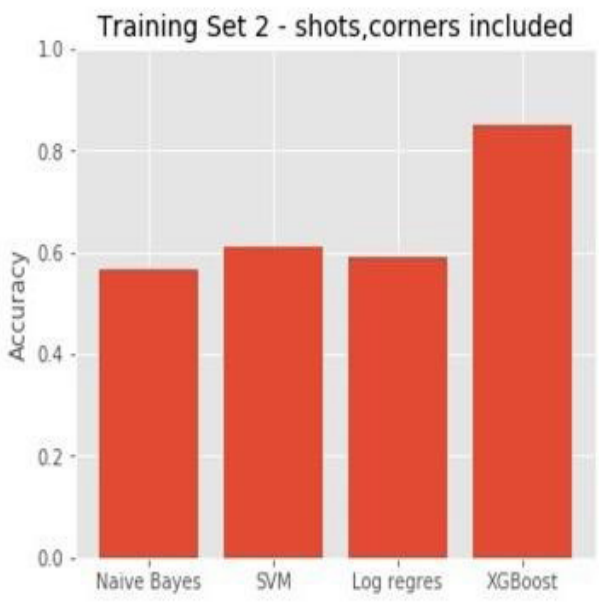

Figure 5: Training set 2

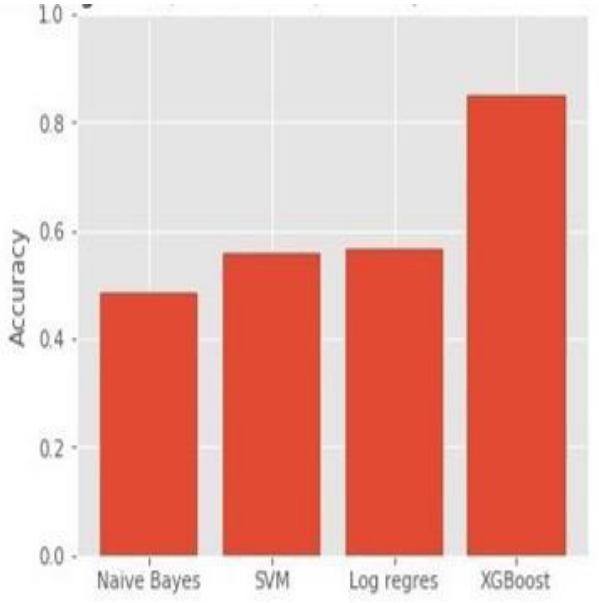

Figure 6: Normalized dataset 1

and data miners collaborate to create the best models for forecasting and explaining results. We have used it to obtain data sets useful for our model Figure 4 and Figure 5.

The performance of Naive Bayes and SVM decreases after normalization, while the performance of Logistics Regression increases after normalization as seen in Figure 6 and Figure 7..

The final score will be 0 and 1 where 0 means the home team loses and 1 means the other (Draw or Away side wins).

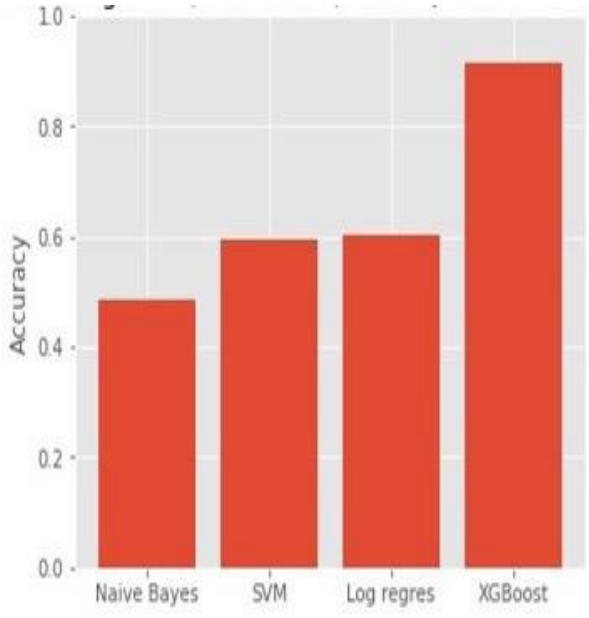

Figure 7: Normalized dataset 2

6. CONCLUSION

Today, football is one of the world's most popular sports and has a large base of fans. The outcome of a football match is hard to predict, as too many factors can be taken into account. There are several places where changes will lead to improved accuracy as no model is flawless. This may be in the form of bigger sets that can be obtained by future matches. Further, matches will require further sets of data. Besides, this will lead to more and more functional feature sets.

6.1. Usage and Future Work

We maximized the characteristics of the teams in the function vector because it is feasible to use our model to handle the football clubs and their personnel. If we consider the current player situation, customization based on players can be done even before the match begins. The analysis of matches for gaming purposes would be a more usage case. In this case, we would recommend that the betting odds and few other features for the specific match are also included in the model. This research may prove to be a good turning point in terms of start-ups. This model is then used to create an app or website for wagering. The model can also be implemented in a website which will help in journalism and the league as well. Twitter sentiment analysis: Live data will be fetched from Twitter by analyzing the sentiment of the posts posted on the platform. If a majority of these posts reflect positivism, the corresponding team is likely at a winning stage [6]. Conversely, if the posts are negative, it could mean that the team is playing poorly and it is likely that it would lose the match.

7. ACKNOWLEDGEMENT

We would like to thank Ms. Nida Parkar, for her full support and guidance. All of her contributions have found a place in our paper. During the implementation process, the technical guidance that we received under her care proved to be very helpful. From the very first day, she was a source of inspiration for us. It was a memorable experience to benefit from her advice. We would also like to thank Ms. Suvarna Pansambal and all the other staff members of the Computer Department who have guided us through this paper idea. Ultimately, we want to thank all our teachers and friends who helped and led us in our studies.

8. REFERENCES

[1] Baio G., & Blangiardo M. (2010, January 21). Bayesian hierarchical model for the prediction of football results. Taylor & Francis. https://www.tandfonline.com/doi/abs/10.1080/02664760802684177

[2] Hucaljuk, J. & Rakipović, A. (2011). Predicting football scores using machine learning techniques - IEEE conference publication. IEEE Xplore. https://ieeexplore.ieee.org/document/5967321

[3] Tax, N. & Jousts, Y. Predicting the Dutch football competition using public data: A machine learning approach. (2015). Academia.edu - Share research. https://www.academia.edu/16272629/Predicting_The_Dutch_Football_Competition_Using_Public_Data_A_Machine_Learning_Approach

[4] Baboota, R., & Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. ScienceDirect.com | Science, health and medical journals, full text articles and books. https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116

[5] Khazaal, Khazaal, Y., Chatton, A., & Billieux, J. (2012). Effects of expertise on football betting.Semantic Scholarhttps://www.semanticscholar.org/paper/Effects-of-expertise-on-football-betting-Khazaal-Chatton/fd130aa25dcc3d2fe4771ad9d36ba7f1c612a4d0 .

[6] Kampakis, S., & Andreas, A. (2014, November 5). Using Twitter to predict football outcomes. ResearchGate.https://www.researchgate.net/publication/267869775_Using_Twitter_to_predict_football_outcomes

[7] Peng, J., Lee, K., & Ingersoll, G. (2002).An Introduction to Logistic Regression Analysis andReporting. Taylor & Francis. https://www.tandfonline.com/doi/abs/10.1080/00220670209598786

[8] Peng, J., Lee, K., & Ingersoll, G. (2002). An Introduction to Logistic Regression Analysis and Reporting. Taylor & Francis. https://www.tandfonline.com/doi/abs/10.1080/00220670209598786 N. Ancona, G. Cicirelli, A. Branca and A. Distante, "Goal detection in football by using support vector machines for classification," IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001, pp. 611-616 vol.1.

[9] Yang, F. (2018). An Implementation of Naive Bayes Classifier. Semantic Scholar. https://www.semanticscholar.org/paper/An-Implementation-of-Naive-Bayes-Classifier-Yang/2e01513a6acff9875b44c2c59c939e2ca50c3162

AUTHORS

Wasim Gourh
Student Atharva College of Engineering

Keshav Poojary
Student Atharva College of Engineering

Mallika Vengarai
Student Atharva College of Engineering

Nida Parkar
Assistant Professor
Atharva College of Engineering

