Full Time Result Prediction using Ensemble Techniques

Mrigank Vashist
Student, Department of Information Technology
Maharaja Agrasen Institute of Technology
IP University
New Delhi, India
mrigankvashist@gmail.com

Vasudha Bahl
Assistant Professor, Department of Information Technology
Maharaja Agrasen Institute of Technology
IP University
New Delhi, India
vasudhabahl@mait.ac.in

Dr. Amita Goel
Professor, Department of Information Technology
Maharaja Agrasen Institute of Technology
IP University
New Delhi, India
amitag@it.mait.ac.in

Nidhi Sengar
Assistant Professor, Department of Information Technology
Maharaja Agrasen Institute of Technology
IP University
New Delhi, India
nidhisengar@mait.ac.in

***Abstract* - Sports Analytics is a growing industry and one of the best real-word applications of Data Science. In this paper, the interest of author and machine learning capabilities were combined to develop a result predictor for football matches. The model proposed is capable of predicting result of any English Premier League Match at the half-time with 75% accuracy. The full-time result predictor is a system based on ensemble of powerful classification algorithms which can predict the odds of winning and draw of both home team and away team on the basis of goals scored at the half time and the current standings in the league. The model learns from the past records of the league and the results of different models are compared in the last section of the paper.**

Keywords— Data Science, comparative models, result prediction, football analysis

I. INTRODUCTION

Football being a very popular sport around the globe is followed by millions. The interest in the game and the rise of data analytics has led to the collection of a lot of data related to many games in many countries such as the one including information about each goal and even each shot or pass. This data when combined with different analytics has many possible applications such as match analysis, identifying playing styles of the player, player valuation, match tactics, performance prediction, outcome and league table position prediction.

One of the most important element of Machine Learning in football is the capability of evaluating performance of a team and using it further to predict the results of matches in future. The three possible outcomes are: win, loss and draw. This looks like a multi-class classification problem and thus, traditional predictive methods have treated it like one and created statistical models for result prediction using the historical match results. However, football games are low-scoring in nature and there is a randomness related to the goals scored. Also, as every match is different, so predicting the game results before the commencement of future matches will be logically inaccurate.

With the emergence of robust machine learning techniques, the predictive performance of classification problems have been improved over last few years. This paper explores different data preparation techniques and algorithms to predict the result of a football match in terms of odds of winning, losing or drawing. The state-of-the-art is extended by combining the popular prediction methods, team ratings of the attacking and defensive team and the goals scored by them at half time. This has been possible due to the huge dataset available publicly.

The data for the research is of good quality and extracted from credible data sources. There are huge number of statistics available, however, for the ease of accessibility of the system, only a few would be considered for now. The main goal is to build a model capable of understanding team's historic performances and predicting the odds at the half-time of the concerned match. The important aspects of the research are building a training and testing pipeline that can compare the benefits of adding new features and using other models on the result.

The next sections of the paper include background of football and the research work done in this field. Then. the origin of data, its pre-processing, features and modelling. Further, different classification models are trained and compared against the validation set. In the final section of the paper, the results of the comparison and the functioning of the system is described, followed by conclusion and references.

II. BACKGROUND

Football matches generally comprises of two teams- the home team and away team depending on the ground the match is being played on, i.e., if the match is played at team A's stadium, then it would be termed as the home team and the other as away team. Each time has 11 players and the total playing time is 90 minutes, known as the full time. This particular paper focuses on predicting the result at half-time, i.e., after 45 minutes of playing time. After half-time, there is a 20 minutes break before beginning of second half.

The game's scoring metric is a goal. At the end of the match, the team with the highest number of goals win and the other team loses. If both the teams have equal number of goals at the end of the match, then, it is called a draw. Apart from the playing time, there are some penalties and free kicks which can provide the teams to score goals when the other team commits a foul.

Almost every European country has a domestic league where teams, generally 20 plays against each other, each twice in the league as home and away team. The winning team and losing team score 3 points and 0 points respectively. if it's a draw, then 1 point each is given to each team. The team/club with the highest number of points in the end of the league wins the league. In this paper, both the individual goals scored by the teams by half-time and the number of points scored by the teams in the league are considered. On the basis of points scored in the league, every team secures a rank on the leaderboard which is also going to be taken into consideration for the prediction.

Other rules and terms related to football are not required for the purpose of this research.

III. LITERATURE SURVEY

In this section, past research done on the subject is detailed along with the linkage to this research paper. The paper 'Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers of 2021 defined many parameters which can be used to assess a player and the success of a transfer. The research used algorithms like Random Forest, AdaBoost and Naïve Bayes to train and test the classifiers, which are some of the algorithms that would be used in this research also. Numerous experiments have been performed with different parameter weights and this forms a basis of experiments performed in this research.

The paper 'A deep learning framework for football match prediction' by Md. Ashiqur Rahman compared his proposed approach using deep neural model with classical machine learning algorithms based on features. Experimentation was performed for selecting and evaluation features that would help in automatic prediction of football match results. The inspiration of features used in this research is drawn from this paper. Another deep learning-based paper called 'A deep learning framework for football match prediction' by the same researcher excellently predicted the results of FIFA 2018 world cup by dividing data into training, test and validation. The paper narrowed down the research to one single league which is taken as reference for this research on English Premier League data in the similar manner.

Rudraksh Tuwani performed an analysis on Football data and made it accessible to public on his GitHub. In this repository, along with the insights gained using the analysis and the code for building a prediction system are given. This system predicts whether the home team will win the match or not. 3 different models: LogisticRegression, Support Vector Classifer and XGBoost Classifier are trained and parameters are tuned for best performance. The best results were given by XGB Classifier, i.e., an accuracy of 69.23 %. This was taken as a reference for the comparisons and research performed in this paper.

IV. DATASET

A. Origin

For this research, one of the most eminent international football leagues, Premier League or English Premier League was chosen. 20 clubs participate in the league with each team playing 38 matches throughout the season. The first premier league was hosted in 1992 and since then many new English clubs have come up, so, for this research, data from the year 2005 to 2020 was considered.

The data was retrieved from the Football-Data.co.uk website. The website lists the results of five English leagues for every year starting from 2005. Seventeen different datasets were taken from the website such that the results of the year 2005 to 2020 behave as the training data and the ongoing season 2021-2022 behaves as the test dataset. All these datasets have information about full time and half-time results, match stats, total goals and Away/Home odds.

B. Features

There are a large number of features provided. However, in the data-cleaning process, most of them are eliminated and only the most easy-to-comprehend features are used. This is because the end goal of this research is to create an application capable of predicting the results of English Premier League matches which can be used by anyone. So, for such an application, the input from the user needs to be minimum and easily available. Although for improving the model in future, more features would be added as more the data, better the model will recognize patterns. For this research, the following features are considered:

TABLE I. FEATURES OF THE DATASET

Name of the feature	**Data Type**	**Definition**
HomeTeam	Categorical object	Name of the home team
AwayTeam	Categorical object	Name of the away team
HTHG	Integer	Half Time Home Goals
HTAG	Integer	Half Time Away Goals
FTR	Integer	Full Time Result
HomeTeamLP	Categorical Integer	Home Team Leader Position
AwayTeamLP	Categorical Integer	Away Team Leader Position

In this data, HTHG and HTAG are the goals scored by the HomeTeam and AwayTeam respectively at half-time. FTR suggests the result of the match (win, lose or draw) which is the dependent variable in this case. The HomeTeamLP and AwayTeamLP are the leaderboard positions of the both teams at the end of the last season, i.e., if a match is being played in 2021-22 season by team A and team B, then HomeTeamLP and AwayTeamLP would have the rank secured by the teams (between 1 to 20) in the 2020-21 season.

C. Data Pre-processing

Points scored by each team is an important metric in football analysis as the final leaderboard is dependent on it. Thus, points are calculated for each team in the dataset for every match on the results of previous matches as discussed earlier. Thus, two features are added. HTP and ATP are the points scored up to the particular match in the league by the HomeTeam and AwayTeam respectively. This value stays 0 unless the team has played its first match of the season and is entered by the user in the application.

For making the model understand the importance of points and leaderboard positions, two features were added called as DiffPts and DiffLP. The former is the difference of HTP and ATP whereas, the latter is the difference of HomeTeamLP and AwayTeamLP. These values can be positive, negative or zero. For making the application easy-to-use, the values of HomeTeamLP and AwayTeamLP are extracted from a separate database and not inputted by the user. Similarly, the DiffPts and DiffLP are calculated as a part of pipeline and not entered by the user.

As HomeTeam and AwayTeam are categorical variables, for the model to understand this, the variables need to be one-hot encoded. As a result, a 40x40 binary matrix is created where 40 is the number of teams that ever participated in the Premier League from 2005 to 2020, i.e., every year only 20 clubs participate but they might not necessarily be the same. Now, every team is encoded as a unique string of 0s and 1s.

The training dataset has data of 6080 matches, each row representing a match and total 11 features, i.e., 11 columns. Shape of dataset: 6080x11. After one-hot encoding the data, the shape becomes 6080x89 as 40 encoded features for the home team and away team each are concatenated. To feed this data to different models, it is necessary that data lies in a common range. Out of the 88 dependent variables, all variables except the encoded features show high variance. Thus, it is essential to scale these variables. This standardizes the variables by removing the mean and scaling them to unit variance.

V. Data Analysis and Modelling

A. *Exploratory Analysis*

The end task is to predict the winner of the football so first there is a need to dig out factors that might influence the win. The first factor that can largely impact is the stadium or the ground. In the last 15 years, 46% of times a team has won when it is playing on its home ground. Figure 1 shows the aggregate win percentage of the last 15 years of English Premier League.

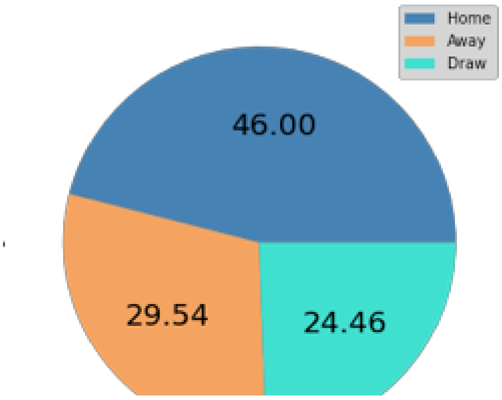


Fig. 1. Distribution of match wins from 2005-2020

Next, number of goals scored by a team in the first-half of the match is a very important deciding factor because if one team certainly takes the lead before half-time, generally the game is theirs. This can be seen in Figure 2, scatterplots between Full-Time Result and goals scored by home team and away team by half-time. The general trend says if a team score more than 2 goals in the first half, they would mostly be the winner. Though nothing is fixed in sports but this is what past statistics suggest.

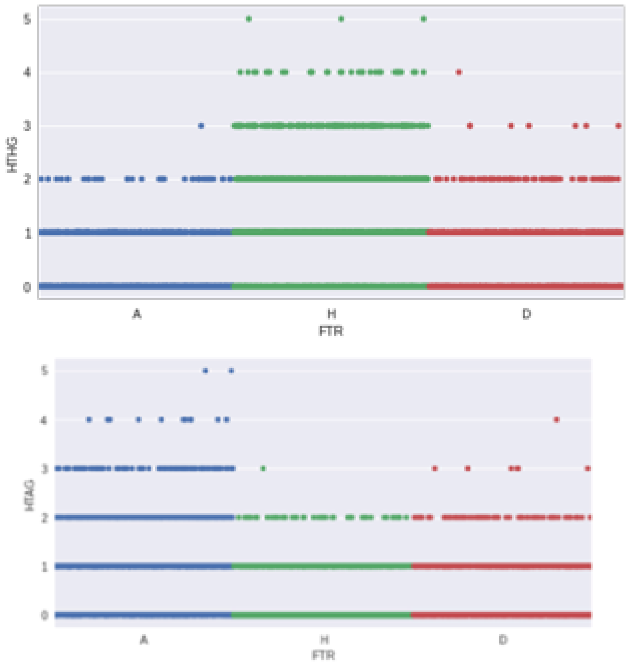


Fig. 2. FTR v/s HTHG and HTAG

As mentioned in the first section, football is a game of a few goals and this can be seen from the line plots in Figure 3. The former suggests the tendency of the home team is to score 1 goal in the first half which leaves with almost equal probabilities of winning, losing and a draw. However, one goal less makes increases the chance of winning of the away team and one goal more increases chances of the home team.

Next important factor is the leaderboard standings. For this research, the leaderboard position of last year is considered. This might not be a very strong determinant because teams recruit new players, coaches and can improve as the season goes on but can help determine but it is highly unlikely for a strong team to lose matches to a new team in the league. The normed histogram in Figure 3 suggests that there is a higher chance of the home team winning a match this season if it finished up to the 10th rank in the last season.

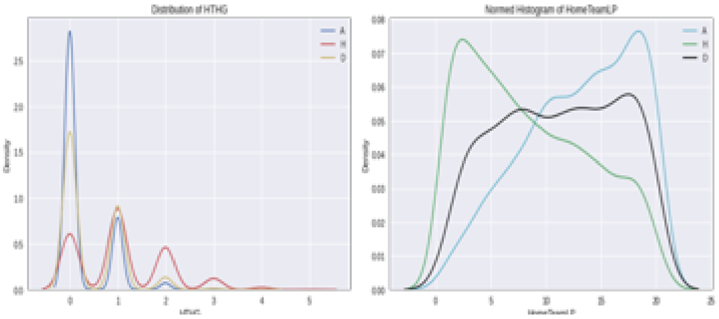


Fig. 3. Density plots for HTHG and HomeTeamLP

Not only the previous season but how the home team is performing in comparison to the away team this season is also an interesting parameter to consider. In Figure 4, 0, 1 and 2 suggests home team win, away team win and draw respectively. This plot suggests that the trend of last 15 years suggest that if home team has 2 points more than the away team, it increased its chances of winning the match and its very rare for away team winning in such a circumstance.

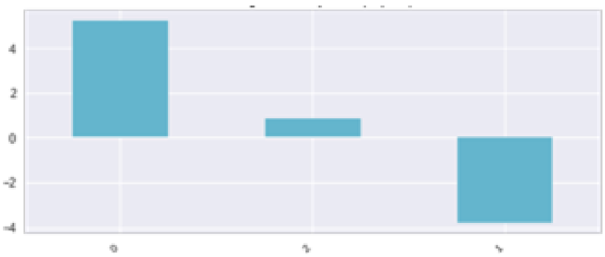


Fig. 4. Average DiffLP by FTR

B. *Modelling and Testing*

The end goal is to predict the probability of winning for both the playing teams and a draw. For this, it is essential to select an algorithm capable of not giving the best classification accuracy and then use that algorithm for predicting the probability using the predict_proba function. As it is a multi-classification problem, the first approach was to use state-of-the art classification algorithms. For this purpose, 9 of the most common classification algorithms namely XGBClassifier, Logistic Regressor, Random Forest Classifier, Decision Tree Classifier, K-Neighbors classifier, Support Vector classifier, Naïve Bayes Classifier, Ada Boost Classifier and Gradient Boosting Classifier were trained and fit to the training data.

For validation, two different methods were considered:

- Random

  The training set in this case consisted of 6060 matches from 2005 to 2020. While, validation set has any 20 matches apart from the training set.

- Sequential

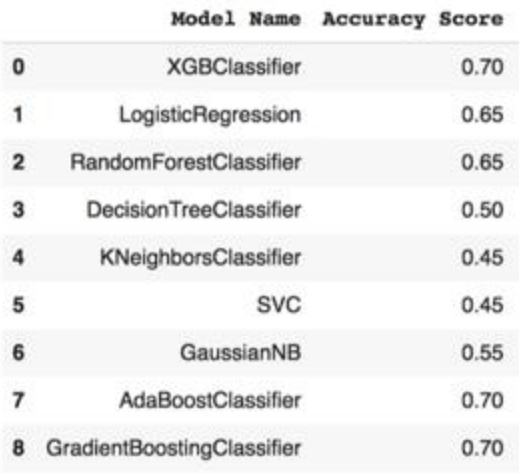
  The training set in this case consisted of 6060 matches from 2005 to 2020 but in order. The validation set in this case consisted of the last 20 matches from the season 2020-2021.

The next approach, i.e the ensemble approach was to combine the most powerful algorithms using Voting Classifier to see the effect on accuracy. For these different combinations of most fruitful algorithms were trained and their results were compared for the sequential approach only given that the aim of the model is to predict the result of future matches which behaves in a sequential manner.

VI. RESULTS

On the basis of the feature importance, exploratory data analysis and best scaling methods suitable for the data, the data was prepared and algorithm was modelled around it. The first approach while modelling was to find out most powerful algorithms and the other one used this knowledge for getting the best accuracy for the use case. The first approach used two methods: the random and sequential one, results of which are presented in Table 1 and Table 2 respectively.

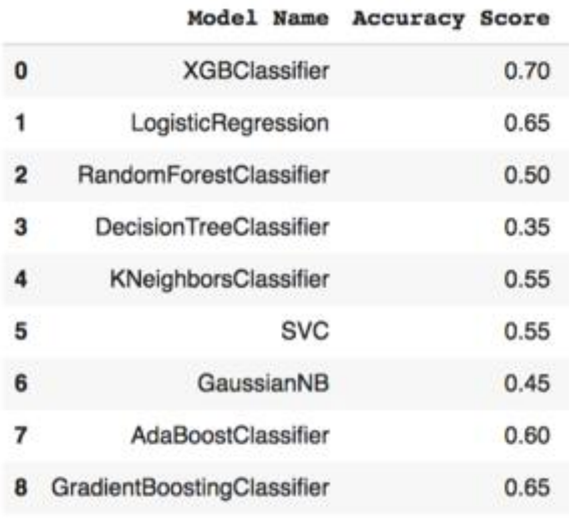
TABLE II. COMPARISON RESULTS OF CLASSIFICATION MODEL-I



	Model Name	Accuracy Score
0	XGBClassifier	0.70
1	LogisticRegression	0.65
2	RandomForestClassifier	0.65
3	DecisionTreeClassifier	0.50
4	KNeighborsClassifier	0.45
5	SVC	0.45
6	GaussianNB	0.55
7	AdaBoostClassifier	0.70
8	GradientBoostingClassifier	0.70

The highest accuracy is delivered by XGB Classifier, AdaBoost Classifer and Gradient Boosting Classifier.

TABLE III. COMPARISON RESULTS OF CLASSIFICATION MODEL-II



	Model Name	Accuracy Score
0	XGBClassifier	0.70
1	LogisticRegression	0.65
2	RandomForestClassifier	0.50
3	DecisionTreeClassifier	0.35
4	KNeighborsClassifier	0.55
5	SVC	0.55
6	GaussianNB	0.45
7	AdaBoostClassifier	0.60
8	GradientBoostingClassifier	0.65

The clear winner from both the cases above is XGBoost Classifier. However, for the next step Logistic Regression and Random Forest Classifier are also taken into consideration.

As per the second approach, the top 5 algorithms were trained for different algorithms and the results are tabulated in Table 3.

TABLE IV. COMPARISON RESULTS OF ENSEMBLE ALGORITHMS

Ensemble Combinations	Accuracy
All 9 algorithms	0.65
XGBoost + AdaBoost + GradientBoosting + RandomForest + LogisticRegression	0.7
XGBoost + AdaBoost + GradientBoosting + LogisticRegression	0.65
XGBoost + AdaBoost + GradientBoosting + RandomForest	0.6
XGBoost + AdaBoost + RandomForest	0.65

XGBoost + AdaBoost + LogisticRegression	0.7
XGBoost + GradientBoosting + RandomForest	0.7
XGBoost + GradientBoosting + LogisticRegression	0.75
AdaBoost + GradientBoosting + RandomForest	0.6
XGBoost + RandomForest + LogisticRegression	0.6
AdaBoost + RandomForest + LogisticRegression	0.6
XGBoost + GradientBoosting + AdaBoost	0.65
XGBoost + AdaBoost	0.55
XGBoost + GradientBoosting	0.65
GradientBoost + AdaBoost	0.55

The best results were observed by the combination of three powerful algorithms: Logistic Regression, XGBoost and GradientBoosting Classifier. This model was further deployed and linked to a web application using Flask. The output of this application is shown in the section.

Here is a comparison of the proposed model with other research work done on English Premier League datasets with goals as multinomial classification:

TABLE V.

Name of the research paper	Model	Accuracy
Premier League Match Result Prediction using Machine Learning by Sushant(2019)	Logistic Regression	65.63%
Predicting the Outcome of English Premier League Matches using Machine Learning by Muntaqim Ahmed Raju(2021)	Proposed model based on 5 algorithms	70.2%
Betting on the English Premier League by Nick Campanelli (2019)	Multinomial Logistic Regression	59.5%
Full Time Result Prediction Using Ensemble Techniques (this research)	XGBoost + GradientBoosting + Logistic Regression	75%

VII. Conclusion

Sports analytics is an interesting yet sparsely explored area of machine learning because of the pre-requisite knowledge of the sport, its rules and key-performance indicators. Thus, the goal was to create a football match result predictor with least input from the user with the best possible accuracy. In this paper, the data of one renowned league was taken into consideration however, the approaches can be extended to any football league, national or international. Here, by inputting only 6 features and implementing 9 state-of-the-art algorithms a satisfactory accuracy has been reached.

Given that sports do not run by numbers but players and playing conditions, expecting a very high accuracy would not be possible. However, there are a huge number of statistical indicators and parameters that are left out from this research. In future, more experimentation will be carried out with extra features like results of previous five matches, shots taken, shots at target, fouls, etc. by the half-time to make the model understand better. Not only this, the research can be improved by using neural networks and pre-trained models. Apart from the winning probability, other predictions can also be made such as expected goals and the goals at full time which would make it a regression problem.

References

[1] https://github.com/RudrakshTuwani/Football-Data-Analysis-and-Prediction

[2] https://github.com/llSourcell/Predicting_Winning_Teams

[3] M. A. Raju, M. S. Mia, M. A. Sayed and M. Riaz Uddin, "Predicting the Outcome of English Premier League Matches using Machine Learning," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-6, doi: 10.1109/STI50764.2020.9350327.

[4] Kundu, Tuhin & Choudhury, Akash & Rai, Sruti. (2021). Predicting English Premier League Matches Using Classification and Regression. 10.1007/978-981-15-5077-5_50.

[5] Ulmer, B., & Fernandez, M. (2014). Predicting Soccer Match Results in the English Premier League.

[6] Ćwiklinski, Bartosz & Giełczyk, Agata & Choraś, Michał. (2021). Who Will Score? A Machine Learning Approach to Supporting Football Team Building and Transfers. Entropy. 23. 90. 10.3390/e23010090.

[7] Herbinet, C., 2018. Predicting Football Results Using Machine Learning Techniques. [online] Imperial College London. Available at: <https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf>

[8] Rahman, M.A. A deep learning framework for football match prediction. SN Appl. Sci. 2, 165 (2020). https://doi.org/10.1007/s42452-019-1821-5

[9] Rana, D., 2019. PREMIER LEAGUE MATCH RESULT PREDICTION USING MACHINE LEARNING. [online] Jaypee University of Information Technology Waknaghat, Solan- 173234, Himachal Pradesh. Available at: http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/22987/1/Premier%20League%20Match%20Result%20Prediction%20Using%20Machine%20Learning.pdf

[10] Yadav A, Sharma A, Gautam A, Bathla G, Jindal R (2017) Predicting English Premier League Results using Machine Learning. J Comput Eng Inf Technol 6:1. doi: 10.4172/2324-9307.1000165

[11] Campanelli, N. (2019, May 22). Betting on the English Premier League. Towards Data Science. https://towardsdatascience.com/betting-on-the-english-premier-league-making-money-with-machine-learning-fb6938760c64