

# Predictive Analysis of English Premier League Using Machine Learning

Shruti Singh<sup>1</sup> Eash Saxena<sup>2</sup>

<sup>1,2</sup>Maharaja Agrasen Institute of Technology, New Delhi, India

**Abstract**— Machine Learning allows us to gain insight into data using which we aim to cover feature extraction for premier league football predictive analysis and perform machine learning to gain insight. The system will be performing our analysis based on our featured dataset and implement multiple classification algorithms such as support vector machine, random forest and naïve bayes.

**Key words:** Machine Learning, Data Mining, Classification Algorithm, Feature Extraction, Support Vector Machines, Random Forest, Naïve Bayes

## I. INTRODUCTION

There are 2.3 billion football fans worldwide and 1.2 billion fans of premier league with every match being broadcasted in around 730 million homes [1] premier league is undoubtedly the most followed football league. Sports analytics have been successfully applied to baseball and basketball however there is a need to find out if machine learning can provide insights into the game adored by billions. We will cover existing solutions in terms of feature selection, models and analyse our results. Our system will classify each season which starts in May and ends in August next year in which each team plays 38 matches from which 19 are played on home field and 19 on away field.

## II. LITERATURE SURVEY

Many attempts have been undertaken to uncover patterns based on data of previous seasons, player performance and match statistics. CS229 Final Project from autumn 2013 by Timmaraju et al. [2] used match stats such as corner kicks and shots of previous matches achieving accuracy of 60% but rather limited scope of parameters for broader classification of data.

Research done by Ben Ulmer and Matthew Fernandez of Stanford University [3] used game day data and current team performance achieving error rates of linear classifier (.48), Random Forest (.50), and SVM (.50).

Nivard van Wijk [4] uses the betting concept predicting winner by proposing two models prediction i.e. toto model and score model. This paper aimed to explain the prediction system mathematically using methods and formulas specified in the article. They obtained accuracy of 53% on their model.

Work of Rue et al. [5], used a Bayesian linear model to predict outcome. They used a time-dependent model taking into account the relative strength of attack and defense of each team.

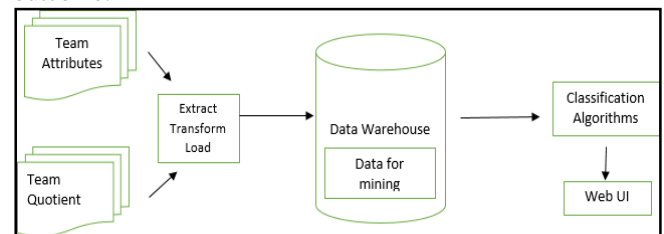
Joseph et al[6] used Bayesian Nets to predict the results of Tottenham Hotspur over the period of 1995-1997. As it relied upon trends from a specific time is was not extendable to later seasons, and they report vast variations in accuracy, ranging between 38% and 59%

The paper on using FIFA game data by Leonardo Cotta et al [7] which compared and contrasted between the Brazilian and German National teams in 2014 and FC Barcelona's distinguished style in the 2012/13 season. This

gave us a new direction to pursue our research leveraging the data of previous seasons with that from Fifa.

## III. SYSTEM OVERVIEW

Features of our system include team attributes which are crawled from the web and computed by taking mean of the cumulative ratings of players and team quotient. These factors are then transformed into a single.csv file. This data for mining is classified into Home, Away or Draw for each individual fixture considering the parameters by various classification algorithms. The outcome in the form of confusion matrix which compares the actual outcome to the predicted outcome which is then displayed on the web user interface. We will now consider the steps in attaining the outcome.



- Team Attributes: Home Team rating and Away Team Rating
- Team Quotient: Home Team Quotient and Away Team Quotient
- Data For Mining: Team Attributes+Team Quotient
- Classification Algorithms: SVM, Random Forest Logistic Regression

## IV. DATASET

We prepared dataset by web crawling of team ratings from so <http://football-data.co.uk/englandm.php>[8] and considering the performance of each team at home field and away team. Our final dataset consists of fifa ratings of each team along with their performances of last 10 seasons[9]. Feature Selection: When dealing with football matches various factors come into play i.e. the playing conditions (home or away), fatigue levels, team selected by the manager and many other factors. Based on our dataset of last 10 years the team which is playing at home has a win percentage of 46.5% away team has a win percentage of 28% and 25% matches end up as draw. Analysing the quality of the team and its opponent is done by taking mean of the player ratings data obtained from <http://football-data.co.uk/englandm.php> thus forming the team rating. We derived home team quotient [10] and away team quotient by using the formula

$$\text{Home Team Quotient} = \frac{\text{Games won by home team}}{\text{Total number of games at home}}$$

$$\text{Away Team Quotient} = \frac{\text{Games won by away team}}{\text{Total number of games away}}$$

This allows us to understand the performance of each team at its home and away ground and take into consideration its associated form along with team ratings.

## V. ALGORITHMS

In this paper, some of the classical machine learning algorithms are used, these algorithms were of great use and helped us in determining the best among them, all them have different test cases and works best with the different dataset and give different accuracy with different dataset but the dataset of this paper was MNIST dataset of handwritten digits, by keeping the data standardised we were able to compare the accuracies with a common metric of accuracy probability.

### A. Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalisation error for forests converges as to a limit as the number of trees in the forest becomes large. The generalisation error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favourably to Ada-boost, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance.

### B. Support Vector Machine

Support Vector Machine (SVM), is applied on different data (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi class. SVM, a powerful machine method developed from statistical learning and has made significant achievement in some field.

SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms

### C. Logistic Regression

Logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical

Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, "0" and "1" (which may represent, for example, "dead" vs. "alive" or "win" vs. "loss").

## VI. ACCURACY COMPARISON

### A. ACCURACY

Algo used	Accuracy
Random Forest	78
SVM	75
Logistic Regression	76

## VII. CONCLUSION AND FUTURE WORK

The comparison of various classification algorithms on various parameters in order to predict the outcome of a

football match has been performed. The best results are obtained when XGBoost is used i.e. an accuracy of 80%. This means every 4 out of 5 matches could be successfully predicted thus resulting in an overall profit for the better.

While the Random Forest provides an accuracy of 78% which further supports the results or predictions made using the features we selected.

## VIII. SOFTMAX

### A. Soft-Max Combination of One-Versus-All Classifiers

Suppose there are  $M$  classes and  $l$  labelled training data  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)$ , where  $\mathbf{x}_i \in \mathbf{R}^m$  is the  $i$ th training example and  $y_i \in \{1, \dots, M\}$  is the class label of  $\mathbf{x}_i$ . For an example  $\mathbf{x}_i$ , let us denote the output (decision function value) of the  $k$ th binary classifier (class  $\omega_k$  versus the rest) as  $r_k^i$ ;  $r_k^i$  is expected to be large if  $\mathbf{x}_i$  is in class  $\omega_k$  and small otherwise.

After  $M$  one-versus-all binary classifiers are constructed, we can obtain the posteriori probabilities through a soft-max function

$$P_k^i = \text{Prob}(\omega_k | \mathbf{x}_i) = \frac{e^{w_k r_k^i + w_{k0}}}{z^i}, \quad (1)$$

where  $z^i = \sum_{k=1}^M e^{w_k r_k^i + w_{k0}}$  is a normalization term that ensures that  $\sum_{k=1}^M P_k^i = 1$ .

1. The parameters of the soft-max function,  $(w_1, w_{10}), \dots, (w_M, w_{M0})$ , can be designed by minimizing a penalized negative log-likelihood (NLL) function, i.e.,

$$\min E = \frac{1}{2} \|\mathbf{w}\|^2 - C \sum_{i=1}^l \log P_{y_i}^i \quad (2)$$

$$\text{subject to } w_k, w_{k0} > 0, k = 1, \dots, M \quad (3)$$

where  $\|\mathbf{w}\|^2 = \sum_{k=1}^M (w_k^2 + w_{k0}^2)$  and  $C$  is a positive regularization parameter.

Note that positiveness constraints are placed on weight factor  $w_k$  and bias factor  $w_{k0}$ . We place positiveness constraints on  $w_k$  because we assume that  $r_k^i$  is large if  $\mathbf{x}_i$  is in class  $\omega_k$  and small otherwise. We place positiveness constraints on  $w_{k0}$  simply to reduce redundancy, since adding a same constant to all  $w_{k0}$  does not change the posteriori probability estimates in (1).

The above constrained optimization problem can be transformed to an unconstrained one by using the following substitute variables

$$s_k = \log(w_k) \text{ and } s_{k0} = \log(w_{k0}), k = 1, \dots, M. \quad (4)$$

The unconstrained optimization problem can be solved using gradient based methods, such as BFGS [3]. The first-order derivatives of  $E$  with respect to the substitute variables can be computed using the following formulas

$$\frac{\partial E}{\partial s_k} = \frac{\partial E}{\partial w_k} \frac{\partial w_k}{\partial s_k} = \left( w_k + C \sum_{y_i=k} (P_k^i - 1) r_k^i + C \sum_{y_i \neq k} P_k^i r_k^i \right) w_k, \quad (5)$$

$$\frac{\partial E}{\partial s_{k0}} = \frac{\partial E}{\partial w_{k0}} \frac{\partial w_{k0}}{\partial s_{k0}} = \left( w_{k0} + C \sum_{y_i=k} (P_k^i - 1) + C \sum_{y_i \neq k} P_k^i \right) w_{k0}. \quad (6)$$

### B. Soft-Max Combination of One-Versus-One-Classification

Following the same idea as in the previous subsection, posteriori probabilities can also be obtained by soft-max combination of one-versus-one binary classifiers. For an example  $\mathbf{x}_i$ , let us denote the outputs of one-versus-one

classifier  $C_{kt}$  as  $r_{kt}^i$ . Obviously we have  $r_{kt}^i = -r_{kt}^i$ . The following soft-max function is used to combine the one-versus-one binary classifiers

$$P_{k=}^i = \text{Prob}(\omega_k | \mathbf{x}_i) = \frac{e^{\sum_{t \neq k} w_{kt} r_{kt}^i + w_{ko}}}{z^i}, \quad (7)$$

where  $z^i = \sum_{k=1}^M e^{\sum_{t \neq k} w_{kt} r_{kt}^i + w_{ko}}$  is a normalization term. The soft-max function parameters can be determined by solving the following optimization problem

$$\min E = \frac{1}{2} \|\mathbf{w}\|^2 - C \sum_{i=1}^l \log P_{y_i}^i \quad (8) \text{ subject to } w_{kt}, w_{ko} > 0, k, t = 1, \dots, M \text{ and } t \neq k \quad (9)$$

where  $\|\mathbf{w}\|^2 = \sum_{k=1}^M (\sum_{t \neq k} w_{kt}^2 + w_{ko}^2)$  and  $C$  is a positive regularization parameter. Note that, as in soft-max combination of one-versus-all classifiers, positiveness constraints are placed on  $w_{kt}$  and  $w_{ko}$  for the same reason.

As before, we can transform the above constrained optimization problem to an unconstrained one by using the following substitute variables

$$s_{kt} = \log(w_{kt}) \quad \text{and} \quad s_{ko} = \log(w_{ko}), \quad k, t = 1, \dots, M \text{ and } t \neq k \quad (10)$$

The first-order derivatives of  $E$  with respect to the substitute variables are

$$\frac{\partial E}{\partial s_{kt}} = \frac{\partial E}{\partial w_{kt}} \frac{\partial w_{kt}}{\partial s_{kt}} = \left( w_{kt} + C \sum_{y_i=k} (P_k^i - 1) r_{kt}^i + C \sum_{y_i \neq k} P_k^i r_{kt}^i \right) w_{kt}, \quad (11)$$

$$\frac{\partial E}{\partial s_{ko}} = \frac{\partial E}{\partial w_{ko}} \frac{\partial w_{ko}}{\partial s_{ko}} = \left( w_{ko} + C \sum_{y_i=k} (P_k^i - 1) + C \sum_{y_i \neq k} P_k^i \right) w_{ko}. \quad (12)$$

The proposed soft-max combination method can be used with any binary classification technique with non-probabilistic outputs. In our numerical study, SVMs are mainly used as the binary classification method.

#### ACKNOWLEDGEMENTS

We would like to thank Ms. Meenu Garg for her guidance and support. She was available to address our queries and lead us in the right direction.

#### REFERENCES

- [1] PREMIER LEAGUE GLOBAL FANBASE <<http://fanresearch.premierleague.com/global-fanbase.aspx>>
- [2] A.S. TIMMARAJU, A. PALNITKAR, & V. KHANNA, GAME ON! PREDICTING ENGLISH PREMIER LEAGUE MATCH OUTCOMES, CS229 STANFORD, 2013.
- [3] BEN ULMER AND MATTHEW FERNANDEZ; PREDICTING SOCCER MATCH RESULTS IN THE ENGLISH PREMIER LEAGUE, CS229 STANFORD, 2014
- [4] NIVARD, W. & MEI, R. D. SOCCER ANALYTICS: PREDICTING THE OF SOCCER MATCHES. (MASTER THESIS: UV UNIVERSITY OF AMSTERDAM), 2012.
- [5] H. RUE AND O. SALVESEN, PREDICTION AND RETROSPECTIVE ANALYSIS OF SOCCER MATCHES IN A LEAGUE. JOURNAL OF THE ROYAL STATISTICAL SOCIETY: SERIES D (THE STATISTICIAN) 49.3 (2000): 399-418.

- [6] JOSEPH, A. E. FENTON, & M. NEIL, PREDICTING FOOTBALL RESULTS USING BAYESIAN NETS AND OTHER MACHINE LEARNING TECHNIQUES. KNOWLEDGE-BASED SYSTEMS 19.7 (2006): 544-553.
- [7] LEONARDO COTTA ET AL: USING FIFA SOCCER VIDEO GAME DATA FOR SOCCER ANALYTICS. LARGE SCALE SPORTS ANALYTICS.
- [8] FIFA RATINGS FOR PLAYERS. <[HTTP://SOFIFA.COM/PLAYERS](http://SOFIFA.COM/PLAYERS)>
- [9] CONSIDERING THE HOME TEAM ADVANTAGE <[HTTPS://WWW.BETTINGPLANET.COM/HOME-FIELD-ADVANTAGE](https://WWW.BETTINGPLANET.COM/HOME-FIELD-ADVANTAGE)>
- [10] "SoccerVista-Football Betting." Web. 11 Dec. 2014. <<http://www.soccervista.com/soccerleaguesorderedbynumberofdraws.php>>
- [11] "FIXTURES AND OUTCOMES OF PREMIER LEAGUE MATCHES" <<http://football-data.co.uk/englandm.php>>
- [12] "MARK LAWRENSEN VS. PINNACLE SPORTS." WEB. 11 DEC. 2014. <<http://www.pinnaclesports.com/en/betting-articles/soccer/marklawrenson-vs-pinnacle-sports>>
- [13] [http://www.ijirccce.com/upload/2017/march/73\\_Predictive.pdf](http://www.ijirccce.com/upload/2017/march/73_Predictive.pdf)