ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Predictive Analysis of Premier League Using Machine Learning

Siddhesh Sathe[1], Darshan Kasat[2], Neha Kulkarni[3], Prof. Rachana Satao[4]

B.E. Students, Dept. of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India[123]

Assistant Professor, Dept. of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India[4]

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.

KEYWORDS: 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

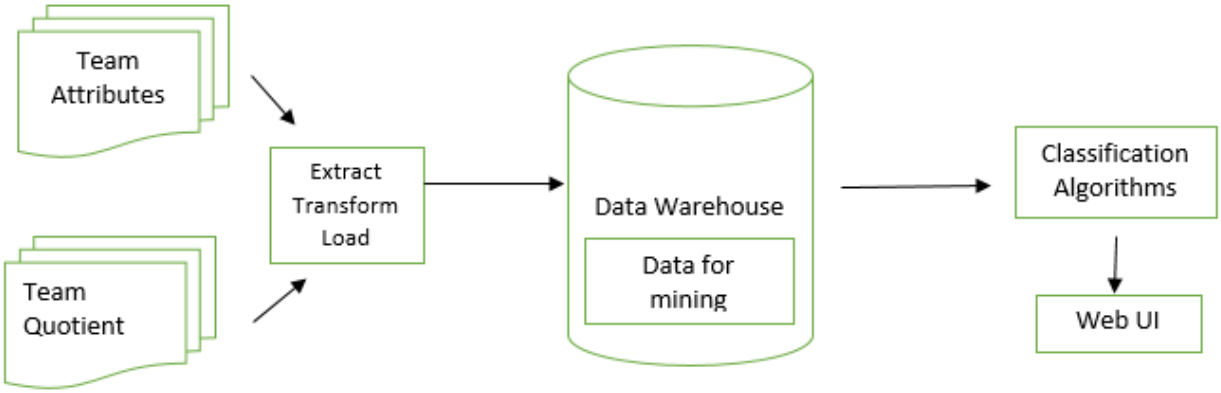
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

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, Naïve Bayes, Random Forest
Web UI: Shiny for R

IV. DATASET

We prepared dataset by web crawling of team ratings from sofifa[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 sofifa 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.

Models: We applied following models for our classification:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Naïve Bayes: A naive Bayes classifier considers features to contribute independently to the probability of the outcome. Naïve Bayes will consider rating for home team, away team and their respective quotient independently to classify full time result. We achieved an error rate of 0.44 for training data and 0.45 for test data.

Support Vector Machine: Support Vector Classifier constructs hyper-planes for classification. SVM kernel RBF has previously been successfully applied previously for classifying football data, it classified our data in three classes Home Win, Away Win and Draw. Best parameters for SVM had cost function of 1, gamma function of 0.2, 995 support vectors achieving error rate of 0.40 for training and 0.42 for test dataset

Random Forest: Random Forest Classification works by collection of non-related decision trees. In our system Random Forest had high error rates compared to SVM and Naïve Bayes of 0.496 for training dataset and 0.498 for test dataset.

V. RESULTS

Results in the form of confusion matrix for our classification algorithms are as follows. Confusion matrix here are for the test data. The result displayed here is predicted outcome versus actual outcome for team playing at Home field.

1. Naïve Bayes:

	Predicted Loss	Predicted Draw	Predicted Win
Actual Loss	106	65	52
Actual Draw	10	15	10
Actual Win	37	84	209

For our test data Naïve Bayes classifier was able to classify 209 wins accurately out of 330 having accuracy of 0.63, 15 draws were accurately classified out of 35 having accuracy of 0.42 and 106 losses out of 223 were classified having accuracy of 0.47.

2. Support Vector Machine:

	Predicted Loss	Predicted Draw	Predicted Win
Actual Loss	90	35	27
Actual Draw	22	46	15
Actual Win	41	83	229

Support vector machine was able to classify 229 wins accurately out of 353 having accuracy of 0.64, 46 draws were accurately classified out of 83 having accuracy of 0.55 and 90 losses out of 152 were classified having accuracy of 0.59

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

3. Random Forest Algorithm:

	Predicted Loss	Predicted Draw	Predicted Win
Actual Loss	69	40	28
Actual Draw	37	39	57
Actual Win	47	85	186

Random Forest was able to classify 186 wins accurately out of 318 having accuracy of 0.58, 39 draws were accurately classified out of 133 having accuracy of 0.29The failure to predict draws accurately is because it is least likely result to occur [11] and 69 losses out of 137 were accurately classified having accuracy of 0.50.

VI. CONCLUSION AND FUTURE WORK

Best performing algorithm in our system was SVM having accuracy of 0.599 followed by Naïve Bayes of 0.55 which is better than accuracy of 0.52 of leading BBC analyst Mark Lawrenson[12] and betting organization Pinnacle Sportsin which had accuracy of 0.55 which is equivalent to that obtained by naïve bayes. Random forest with accuracy of 0.50 had lowest accuracy. This accuracy can be further improved by adding more relevant features developing models which take into consider even broader aspects of football.

VII. ACKNOWLEDGEMENTS

We would like to thank Prof. Rachana Satao for her guidance and support. She was available to address our queries and lead us in the right direction. We would also like to thank Prof. Piyush Sonewar for being around and helping us.

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. A. 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>
9. CONSIDERING THE HOME TEAM 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 LAWRENSON VS. PINNACLE SPORTS." WEB. 11 DEC. 2014. <http://www.pinnaclesports.com/en/betting-articles/soccer/marklawrenson-vs-pinnacle-sports>