A Machine Learning Approach to Analyze the Statistics of Football Players

Debayan Das

PROJECT GUIDE: Prof. M. Varalakshmi

School of Information Technology & Engineering (SITE)
Vellore Institute of Technology (VIT)
Vellore, Tamil Nadu, India

***Abstract*— With the advent of many a machine learning technique and the increasing ease of accessibility of internet connection "Sports data Analysis" has rose as an important domain of research. The main objective of our project is applying the various machine learning techniques on the data of performances of the players of different clubs of Europe's top five leagues and to classify them into three categories i.e. who should be retained by the club, who should be sold and who should be given another chance but must be taken under strict guidance of the coach. Here we have used our own metric to evaluate a player's performance. We have applied two machine learning techniques: SVM and KNN. And by the comparative study of both the models we found that SVM and KNN is giving accuracy 82.03% and 80.67% respectively in order to classify the players. So, we can see that SVM is working better than KNN in this scenario.**

Keywords*— *Football, Machine Learning, SVM, KNN, Python, Webscraping

I. INTRODUCTION

Prediction system is now widely used all over the world in variety of fields such as stock market, crime analytics, online shopping, sports data analysis etc. So it has a big impact on fields where prediction is an integral part of the job. In the field of the sports the system can be used for betting, for a player to improve his game, for coaches to analyze the performance of the squad and enhance the game plan. As a result machine learning approach is a highly trending topic in the field of analyzing the sports data. It will be also helpful for the team management.

In the last few years a lot of work has been done to predict the outcome of the matches. With love and passion for sports and inspiration from many of the currently available applications to analyze the sports data and predict accordingly, we attempt to analyze the performance of players of a team to make the job easier for a coach of a team in order make his squad stronger.

In our model the input of the application is the performance data of football players and outcome of the model will be based on their recent performance what the team management should do with them. Their management can keep them to the team, can sell them or they can give him or her another chance under strict regulation.

II. RELATED WORK

Although it is difficult to decide all features that put an effect on the results of the matches, an attempt is made to find the most significant features and various classifiers are tested to give the solution the problem in the study of the methods of machine learning techniques used in order to predict outcome of soccer matches[1].

In another study the stats of the various clubs of English Premier leagues have been taken and machine learning algorithms were applied on them[2]. These are KNN, SVM, Logistic Regression, Random Forest. They have proposed the outcome of each upcoming match. To do so, they have taken the dataset from a genuine website and then they have extracted the data that they need to train their model. They have used these features directly sometimes and sometimes made their own feature using these to train their ML model.

In some other study they have taken the stats of the various clubs of English Premier leagues and have applied 4 machine learning algorithms on them[3]. These models are KNN, SVM, Logistic Regression, Random Forest. They have proposed the outcome of each upcoming match. In this project, they developed the ’expected goals’ metric which will helped them to evaluate a team’s performance, instead of using the actual number of goals scored. We combined this metric with a calculation of a team’s offensive and defensive ratings which are updated after each game and used to build a classification model predicting the outcomes of future matches, as well as a regression model predicting the scores of future games.

In the study of analyzing performance differences of football players between 2-years prior to signing the new contract and the year after signing a new contract (the following year) while taking playing position, nationality, player’s role, team ability, and age into account the dependent variables studied were: shooting accuracy, defense (the sum of defensive actions, tackles, blocks, and interceptions), yellow cards, red cards, passing accuracy, tackle success, and minutes played per match. The main results (very likely and most likely effects) showed better performance in the year prior to signing a new contract than the previous year for foreign important defenders (decreased number of red cards), national important midfielders (increased number of minutes played), foreign important forwards (increased minutes played and defense), and national important forwards (increased minutes played). In addition, performance was lower the year after signing the contract compared to the previous one for less important defenders (decreasing defense), national less important midfielders (decreased minutes played), and foreign less important forwards (decreased defense).[4]

In the research the stats of the various leagues of 35 countries have been taken and various machine learning algorithm have been applied to predict the outcome of the 206 future soccer matches.[5] To do so, they stored most commonly and freely available as well as consistently reported information about the outcome of a league soccer match. This information concerns the goals scored by each team, teams involved, league, season and the date of the match. To train their ML model, sometimes they have obtained an update of the latest results of teams in the league and sometimes made their own feature, and then predicted the future games.

Another experiment is conducted on by thinking of three different questions.[6] Those are all involving the game prediction and style from a soccer ball-event dataset by the machine learning approach. Here the features were the different ball-event which are pass, shot etc. Here they used three methods like K-NN classifier, SVM and logistic classification methods. The target was the prediction of winning team, sequential passing, tackling and shot accuracy. Main target is the passing accuracy. For soccer using machine learning technique one need to answer several questions to train the model. The features are the number of shots, possession, passing accuracy, number of crosses etc. for three predictions there are three types of features. Sometimes they have created their own features. They achieved 0.84 accuracy rate by logistic regression method, accuracy rate 0.345 by RBF SVM for 20-teams classifiers and accuracy rate of 0.735 by learned K-NN classifier

Another experiment is conducted to find the accuracy of predicting a result by comparing regression and neural models to analyze the statistical data of javelin thrower.[7] After taking the data they analyze the data by Shipro-Wilk Normality Test and Homoginity test. They used two models for that such as nonlinear model and perceptron network. At the end they concluded that the neural model predicted higher quality prediction than the nonlinear regression model. This is mainly done on the statistical analysis of the dataset of the javelin throwers.

In some other analysis of matches of Dutch Football Team and tried to give prediction of winner from that analyzation.[8] They used good data mining technique and machine learning algorithm to predict the winner name. For doing this prediction they chose three models named Naïve bayes Model, K-Nearest Neighbor and Random Tree Model more particularly Generalized Boosted Model(GBM). After using those models one can easily conclude which variable can predict the result more accurately and which can predict less accurately. For training the model they chose the features as game type, squad attribute, players’ attribute and the current form of team. They have added the other more dependent features and train the models for getting the best prediction. After that they compared the prediction results and concluded about the best model among those three.

In the implementation of a frame work for sports data, prediction using machine learning technique. This provides a literature survey in Machine Learning Approach focusing on the application of ANN-Artificial Neural Network.[9] They did this based on some modules such as domain and data understanding, data preparation and feature extraction and then processed the data to build a model. Then they tested that model over a test dataset and deployed it. Their features are known prior to the upcoming matches to be played such as how much break is given between two games, how much they travelled for playing that match. For this they also needed the quality of the game they played, that means how many passes they have played, how much shots hits on target etc. These match related features along with external features is used to train dataset. This is their proposed framework named as 'SRP-CRISP-DM' framework.

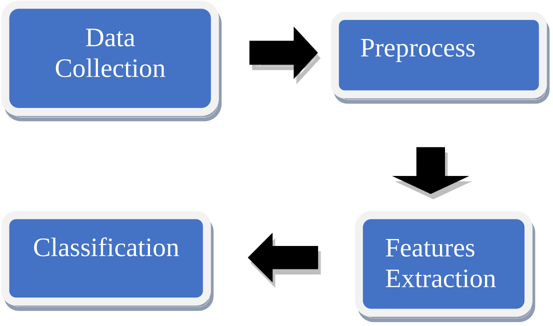
In the test performed on the basis of the factors which are associated for scoring the goals and the scoring opportunities in professional Soccer. Data was collected from 1788 attempts and 169 tries for a FA season.[10] It has been seen that 70% goals are scored from penalty area. So, they used binary logistic regression method over identified 3 covariates which had a significant ($p<0.05$) impact on goals scored. They also added features such as type of shoot, position of goalkeeper and the position of the attempts. Based on these modelled those attempts. The high contribution of factors associated with transitions in played games helped to uncover that importance of tracking goals and goal scoring opportunities back to their point of origin.

III. PROPOSED WORK

Our objective was to design and plan the classification of the football players to decide whether they should be retained by their team or not, through machine learning techniques using python. The proposed strategy includes data collection phase, Feature extracting and metric building stage, applying machine learning techniques on the metric and finally approaching to the classification stage. In the Data collection phase we have collected the seasonwise statistics and market value of the football players of Europe's top 5 leagues and also collected the transfer data. In the feature extraction phase we have collected the features that we needed from the collected data and made the metric by combining the features. In the next phase we have applied the machine learning techniques SVM and KNN on our dataset. And at the end phase i.e. the classification phase we found the classified result with yielded a well defined stratification.

1. Block Diagram

The flow of work for our design is shown below using a block diagram.



A. Data Collection:

The data is collected from the two websites ***whoscored.com*** and ***transfermarket.co.uk*** by webscraping method and stored into excel sheets. We have collected the Players' Performance Statistics, their market values and the transfer details of the teams. Players' performance statistics, market value records, transfer records for both buying and selling of the players. The records are being collected for last 5 years of Europe's top 5 leagues.

B. Pre process

After collecting the data the first step to be done is pre-process the data. As the data is collected for 5 different leagues, those are combined in a single dataset. After combining there are four different datasets: *players performance datasets, market value details, selling details, buying details*. From the Players statistics dataset 3 attributes and 'players name', 'minutes played', 'rating' has been picked up. From the market value dataset we collected 2 attributes 'players name', 'market value' has been picked up. From the buying details dataset 3 attributes 'players name', 'transfer fee', 'market value' has been picked up. From the Selling details dataset only the 'players name' attributes have been picked up. The 'players name' attribute of the players statistics dataset was a multivalued attribute. Keeping only the player's name rest of the data has been stripped off. The price column of both the buying and selling details dataset and the 'market value' column of the Market details dataset contained the Euro sign and units such as 'm' and 'k' they have been stripped off. Apart from that many values for the above attributes contained values like 'Free Transfer', '-', '?'. All these have been replaced by zeros.

C. Feature extraction

The Players statistics dataset and the buying details table has been merged on attribute ‘players name’. Now we have the buyout clause of the players that are bought by the clubs in that season. For rest of the players it contains null value now. All those values are replaced by zero. Now the resulted dataset is joined with the ‘market value’ dataset on the ‘players name’. Now some of the entries in the resulted dataset has market value and rest has null value for that attribute. Here we have used a linear regression technique to predict the market value for the entries that contain null based data of the rows that have valid market values and stored in a new dataset named ‘dataset1’. Now this dataset1 is joined with the Selling Details dataset by keeping only those player names that are in the selling dataset. Now an extra column has been added to that dataset named ‘Class’ with value 0. This dataset contains the players that are sold with class value 0. Now this resulting dataset is joined with the dataset1. So, some rows have the new dataset has ‘Class’ value 0 and rest has null value for that. Now these null values are set as 1. 1 signifies that the players have been retained by their team. Now the ‘Class’ value for the players with rating below ‘6.5’ has been set as 2. 2 signifies the player has not performed well this year but team has kept them in the team to give them another chance. So, the final dataset is consisted of the features ‘Playing Time’, ‘Rating’, ‘Buyout Clause’, ‘Market Value’ and the target variable ‘Class’.

D. Classification

In this paper we have used machine learning techniques to classify whether a player should be retained by his team or should be sold based on the features we have extracted. First the dataset is divided in two parts using *train_test_split* method available in *sklearn* library of python in 80 to 20 ratio. 80% of the data is used to train the models and 20% for testing purpose. Here two supervised machine learning algorithms have been used for the classification purpose. The prepared SVM and KNN method of sklearn library of python has been used here. The model is trained with the training dataset, then it is evaluated with the testing dataset for checking the accuracy of the metric. Then attributes of a player is passed to get his class. And the models predicts the class for the player.

IV Experimentation and Results

A. Dataset Description

The final dataset is a CSV file containing1472 rows. It gives information about the players. The description of the attributes are given in table 1:

Attributes	Description
Mins	The playing time of the players
Rating	The performance of the player out of 10
Market Value	The market value of the player
Fee	The buyout clause of the player
Class	The class of the player

Table 1

For our model the features are *mins, rating, market values* and *fee*. And the target is *class.*

Demonstration of one sample is shown below:

Mins	Rating	Market Value	Fee	Class
1571	6.59	21.6	18	0
1815	7.06	1.35	1.26	1
1441	6.49	27	27.9	2

B. Classification

The data given to the models can be classified into 3 classes: 0, 1, 2. These are described in table 2.

Class	Descrtiption
0	The player should be sold
1	The player should be retained
2	The player should be given another chance

Table 2

For a given value the predction of both the models are showed in table 3.

Model	Mins	Rating	Market Value	Fee	Class
SVM	1416	6.49	1.350000	2.25	1
KNN	1416	6.49	1.350000	2.25	1

Table 3

Fig.1 and Fig.2 gives the plot of original class vs predicted class for both the models SVM and KNN respectively.

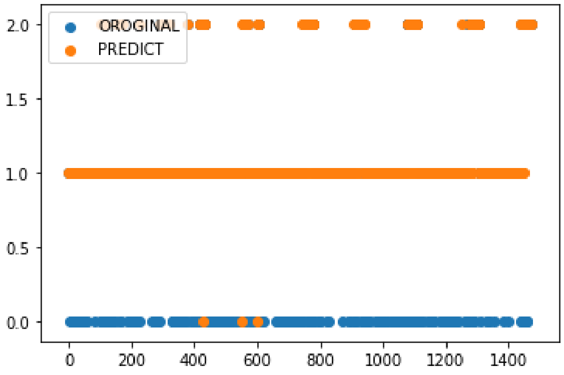


Fig.1

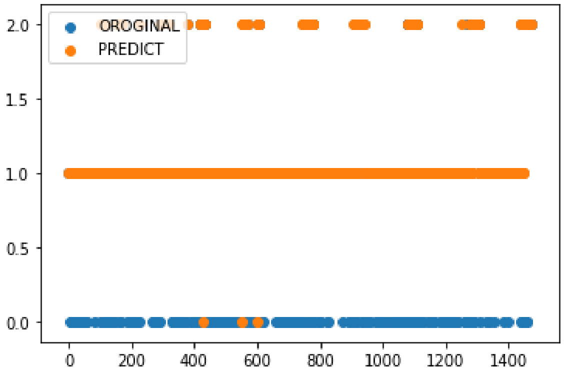


Fig.2

*The overlapping points signify the errors in predicting.

C. Comparison & Analysis:

Table 4 gives the model score of both the models:

Model	Model Score
SVM	0.8203389830508474
KNN	0.8067796610169492

Table 4

By comparing both the algorithms we got that SVM gives better accuracy than KNN

We also found that on the basis of our dataset, SVM model gives the best accuracy with ‘Linear’ Kernel, regularization value(C) 1.0, gamma value 0.01.

Also based on our dataset KNN model gives the best accuracy with ‘Hamming’ distance and K-value=100.

V. CONCLUSION

Though the aim of the project was to evaluate how the machine learning algorithms work on the sport dataset, there are other applications too, such as predicting match outcome, predicting the scores etc. The approach will be helpful for the coaches to make more firm decisions about the team and make the squad stronger and perform in a better way. However, precision is the factor upon which someone could infer and create new knowledge on how to making more firm decisions. We will be glad if our approach really helps someone. Though we proposed the model to be applied on the data of last five years for now we have implemented the model on the data of one year and we have now worked on the four features of the five that we have proposed. In future projects we will increase the size of the dataset and will focus on to increase the related features.

VI. REFERENCES

1. Predicting outcome of soccer matches using machine learning

Saint-Petersburg State University, Mathematics and Mechanics Faculty

2. ENGLISH FOOTBALL PREDICTION USING MACHINE LEARNING CLASSIFIERS

Department of Computer Science and Engineering, SRM IST

3. Predicting Football Results Using Machine Learning Techniques

DEPARTMENT OF COMPUTING, IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

4. Analysis of elite soccer players’ performance before and after signing a new contract Miguel-Ángel Gómez, Carlos Lago, María-Teresa Gómez, Philip Furley
5. The Open International Soccer Database for Machine Learning

Werner Dubitzky · Philippe Lopes · Jesse Davis · Daniel Berrar

6. Machine learning to event data in soccer:

Mathew G.S. Kerr

7. Neural and regression models in sports result prediction:

Adam Maszczyk, Artur Golas et al.

8. Dutch football prediction:

Abel Hijmans in the supervision of Dr. S. Bhulai

9. Machine Learning Framework for Sports Data Prediction:

Rory P. Bunker and Fadi Thabtah

10. Factors Associated with Goals and Goal Scoring opportunities in professional soccer:

Craig Wright, Steve Atkins, Remco Polman, Bryan Jones and Lee Sargeson.