**PREDICTING THE SCORES OF THE FOOTBALL PLAYERS USING AIML**

Aayushman Sharma, school of computer science and engineering, Galgotias University, greater Noida.

Tanu Kumari, School of Computer Science and Engineering, Galgotias University, greater Noida

Dheeraj Kumar swain, school of computer science and engineering, Galgotias University , greater Noida

Gurjas Kaur, school of computer science and engineering, Galgotias University, greater Noida

ABSTRATION

Machine learning is playing an important role in modern technologies. It is an application of artificial intelligence as we know artificial intelligence is how systems can learn automatically and develop without explicit planning. In this project we will be using football players data to predict their overall performance.

Data will include their past performances and their different types of attributes. This project is based on a supervised machine learning problem more specifically Regression Problem i.e. the problem ask to protect the value.

KEYWORDS

Explicit planning, supervised machine learning

INTRODUCTION

Game analysis is an important tool for understanding and enhancing performance in team sports by evaluating the factors associated with success.[01] For its development, different tools are used to describe the statistical events that occur during high-level sports competitions. They allow the identification of tactical, technical and physical features related to the success or failure of a team.

Football statistics have evolved notoriously in recent years due to automatic or semi-automatic detection technologies that provide high-fidelity data streams for each match based on video recordings or observations made with various types of fixed and mobile sensors. Two kinds of data are distinguished. On the one hand, eventing data capture every action related to the ball and its location that takes place during the match. On the other hand, tracking data represent the spatiotemporal position of each player and the ball on the field. However, this growing volume of data has been exploited in a limited way so far. Data science has emerged as a strategic area that, supported by the great possibility of data production for analysis, allows knowledge discovery in sport science with the aim of filling some gaps that traditional statistical methods could not achieve. As a hybrid knowledge area, data science is more than the combination of statistics and computer science, as it requires training in how to interweave statistical and computational techniques in a broader framework, problem by problem, and to address discipline-specific issues.[02]Automated data analysis can help professionals work on optimizing training and competition strategies.

The first sports in which AI analysis of a game began were football and hockey . Research has focused on predicting outcomes relative to previous game statistics, creating models with high accuracy capabilities. Basketball follows a similar path to American football in terms of AI and machine learning, although a wider range of different algorithms, such as logistic regression, support vector machines, neural networks and naı¨ve Bayes, are applied.

LITERATURE SURVEY/ METHODOLOGY

The methodology used in this experiment is based on supervised machine learning. That is, the algorithm ‘‘learns’’ from samples of data to infer a model, and then the model is tested using other samples that were not used to build the model. These test samples allow us to compare the values predicted by the model with actual values, and gauge the model’s accuracy in predicting real samples. In this study, the expected values are the market value of football players, and each sample is represented by a set of variables that represent the player’s performance and skills. The proposed methodology consists of the following steps:

• Step 1: General investigation to determine the factors affecting the players market value: In this step, we have reviewed the studies devoted to predicting a player’s market value. Besides, we searched the literature for the factors affecting the market value, and we identify the variables that have an impact on the market value of players. Nine variables were identified due to their frequent appearance in the literature. Table 1 shows the features that were used in this study.

• Step 2: Pre-processing techniques: Pre-processing is one of the most data mining tasks which includes preparation and transformation of data into a suitable form to mining procedure. It includes several techniques like data cleaning, transformation, reduction, etc [03. Thus, the data has been cleaned and processed, and the most appropriate part of the data was used while building the models. The data cleaning steps are summarized as follows: - Removing the redundant columns (example- name, link, id etc.) and keeping columns with the features we want for modelling as dependent variables (including Age, Height, Potential, International reputation, Weak foot, Team position, Shooting, Passing, and Dribbling). - Dealing with the missing values. - Converting categorical features (Team Position) to numeric values.

• Step 3: The preliminary analysis of a selected subset of features To study the quality of the selected subset of features (identified in step 1), the level of interdependence of these features with each other was studied using the Pearson Correlation Coefficient. The hypothesis on which the heuristic is base states below: Good feature subsets contain features highly correlated (predictive of) with the class, yet uncorrelated with (not predictive of) each other. Figure 1 summarizes the numerical features correlations to the target variable (Player value). Besides this, the predictors (features) themselves were also correlated with each other.

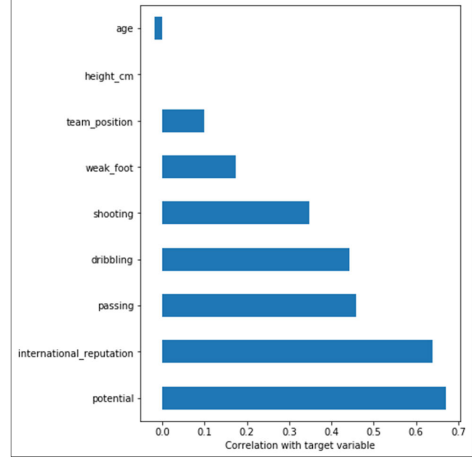


FIGURE 1 - The numerical features correlation to the target variable (Player value).

• Step 4: The extended analysis of a selected subset of features After defining the new subset of features (identified in Step 3), the logical selection of these variables was verified by statistical significance using linear regression analysis and decision trees. In the experiments, an ordinary least squares (OLS) model was fitted. Analysis of variance could provide us with a first impression of how each predictor was correlated with the dependent variable. The P-value was used to determine the statistical significance of the regression coefficients. P-value allows telling whether the null hypothesis is to be rejected or not.

• Step 5: Data splitting After cleaning the dataset and defining the new subset of features and verifying the logical selection of these variables, 80% of the data has randomly allocated to train the classifier, and the remaining 20% was used for testing.

• Step 6: Modelling the market value models In the modelling process, the player market value was used as an objective function of 17980 players. We estimated players’ market values using four regression models that were tested on the full set of features (linear regression, multiple linear regression, decision trees, and random forests).

All models were created using the default parameters unless otherwise noted. The results of the models were compared to the real values, and it is found that it is applicable for this aim.

Step 7: Evaluation models: To evaluate the performance of the models, several metrics were calculated like, Train and Test Split is used to estimate model performance using the training set. Mean absolute errors (MAE), Root mean square errors (RMSE) and The coefficient of determination (R2 ) is used to evaluate the regression models using the testing data. a Python module called scikit-learn is used to build machine learning models.

PROPOSED SYSTEM

To evaluate the performance of the models, several metrics were calculated. For example, the Train and Test Split is used to estimate model performance using the training set. Mean absolute errors (MAE), Root mean square errors (RMSE), and the coefficient of determination (R2 ) is used to evaluate the regression models using the testing data. In machine learning, player market value can be handled in different ways. We can consider it a regression problem and expect the market value based on the data of players’ performance. In this study, we have established four regression models. The data of players’ performance and skills was used as features in building models—to build the baseline and compare results.

RANDOM FOREST REGRESSION

Random forest is an ensemble of decision trees. Many trees, constructed in a certain ‘random’ way, form a Random Forest. Random Forest Regression also provided a significant improvement over the baseline model, with an RMSE of 1.64 and an R-squared score of 0.95. This means that the Random Forest Regression is a powerful generalization of the linear regression algorithm and for Regression trees. Random Forest Regression, like Regression trees, also does a better job at capturing nonlinearity in data by dividing the space into smaller subspaces, depending on the questions asked. As illustrated in Figure 2, each tree is created from a different sample of rows and at each node; a different sample of features is selected for splitting. Then, each of the trees makes its prediction. Finally, these predictions are then averaged to produce a single result. Figure 3 shows the mean absolute error between the predicted and actual values, when using the different machine learning algorithms.[4] As the figure shows, the Random Forest algorithm provided the lowest mean absolute difference

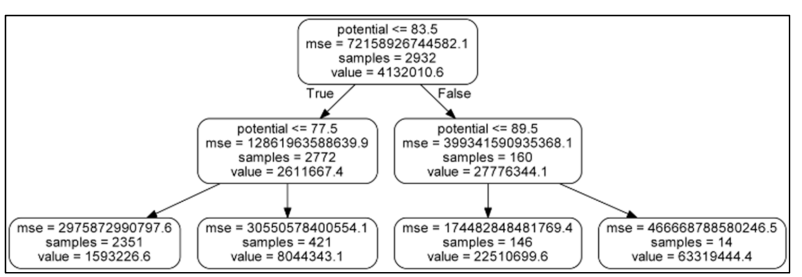


Figure 2 - Random Forest Regression flowchart diagram used for player market value prediction with the number of estimators = 10 and max depth = 3.

between the predicted and actual value, Linear Regression provided the highest mean absolute error.

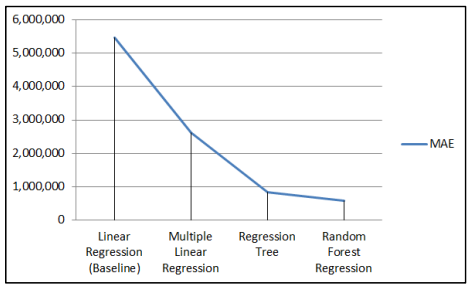


FIGURE 3- . Mean absolute error when using Train and test split scheme.

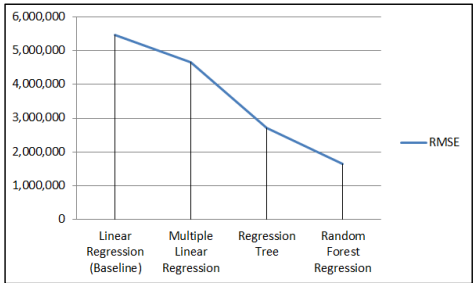


Figure 4- Root mean square errors when using Train and test split scheme.

between the predicted and actual value, Linear Regression provided the highest mean absolute error.

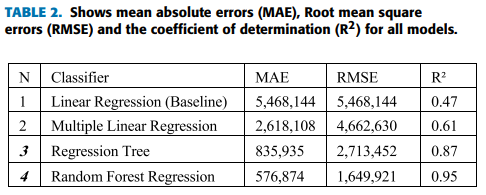


Figure 4 shows the root mean square errors between the predicted and actual values when using the different machine learning algorithms. The Random Forest algorithm provided the lowest root mean square errors difference between the predicted and actual value; Linear Regression provided the highest root mean square errors. Figure 5 shows the coefficient of determination (R2) for each machine learning algorithm used. A value close to 1 indicates a model with close to zero error. In contrast, a value close to zero indicates a model very close to the baseline. As the figure shows, the Random Forest algorithm provided the highest value for the coefficient of determination (R2), Linear Regression provided the lowest value for the coefficient of determination (R2). This means that the Random Forest algorithm is the best for modelling.

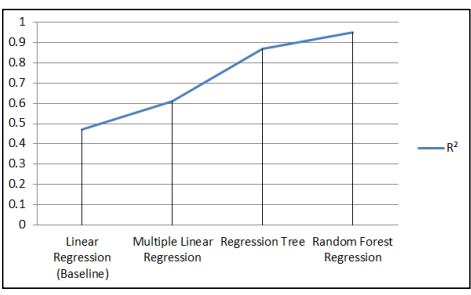


Figure 5- . Coefficient of determination (R2) for each machine learning algorithm.

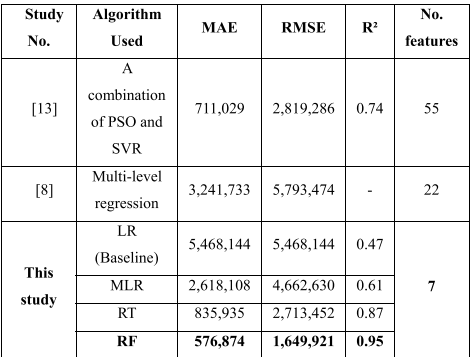


TABLE2 -Evaluation of the studies in the literature

CONCLUSIONS

As a summary of what was presented, the application of AI in the present study structured the positions of the players by their technical-tactical behaviour statistics alone, contributing interesting findings that enrich the understanding of the behaviour profiles and game patterns of the different game positions. This study shows that through the application of AI, it is possible to characterize the positions of players according to their technical-tactical behaviour, exclusively using game statistics from different seasons and national leagues. Its application in the work of club youth teams to detect and identify players with certain qualities can optimize their placement on the field and lead to the improvement of their progression and performance. In addition, in formation ages would help coaches to better identify the positions for players based on their statistics and define which skills are most recommended for the different playing positions. Based on our knowledge, this is the first study of its kind to focus on football, so it is important to continue developing it so that the contributions to the scientific community are conclusive. Moreover, it may be interesting to apply the procedure proposed in this work to

REFERENCES

1. Carling C, Reilly T and Williams AM. Performance assessment for field sports. London: Routledge, 2008
2. Blei DM and Smyth P. Science and data science. Proc Natl Acad Sci USA 2017; 114: 8689–8692.
3. S. A. Alasadi and W. S. Bhaya, ‘‘Review of data preprocessing techniques in data mining,’’ J. Eng. Appl. Sci., vol. 12, no. 16, pp. 4102–4107, 2017.
4. MUSTAFA A. AL-ASADI AND SAKIR TASDEMİR Department of Computer Engineering, Faculty of Technology, Selçuk University, 42130 Konya, Turkey Corresponding author: Mustafa A. Al-Asadi (masadi@lisansustu.selcuk.edu.tr)