NAC Breda's Player Analysis

Vladislav Stoimenov, 235030

ADSAI,

Breda University of Applied Sciences

DISCOVER YOUR WORLD

Friday 26 January 2024

Index

1 Introduction

In this evolving landscape of professional football, data analytics and machine learning are becoming extremely pivotal for enhancing the performance of each team. These methods are very useful for clubs to make informed decisions based on machine learning methods and Exploratory Data Analysis. The research is aimed to help NAC Breda make better decisions on buying new players and strategies in the future to achieve better performance. In football having just a good striker, defender, goalkeeper, etc. cannot be mandatory to perform the best and most of the times would not be enough to achieve team's best efficiency. The clubs have to perform deep research on the player's statistics and how they are performing in different teams and with different players. With the data-driven methodologies, clubs are trying to recognize player's weaknesses and strengths to make strategic decisions that can influence matches outcomes. Here the machine learning part comes, with dataset of players that contains most of their statistics, every team is able to do that.

NAC Breda is striving to improve their team in each possible aspect to make better performance that can lead to more club achievements. This research is mostly aimed to find the best attacking player, that can easily fit the club. It also aims to find the best correlated features from the dataset, on which to find the most suitable one with the greatest development potential

2 Exploratory Data Analysis

The research is based on a given dataset containing more than 16 000 players from different leagues across Europe. The dataset is given by NAC Breda for the research. Each of these 16 000 players contains 114 columns with different data types which represent their personal statistics as a footballer. It contains personal data as their Age, Date of birth, Height, Weight, Club, Positions, Market value, etc. up to their football stats in 3 types (per season, per match (per 90), expected results (xG, xA, etc.)). Most of their statistics are in numerical data, but there are some in categorical, such as their names, positions and others.

The first step of preparing the data for analysis is to identify the missing values. There were a lot of missing values in the dataset. Handling the missing values was done with a for loop to go around all the columns and replaces the numerical data columns with their means and categorical data columns with 'unknown'. According to NAC Players, the one with the highest value costs 850 thousand euros, so the data is filtered for players with market value of under 1 million. To make it easier Market_Value_bin column was created which categorizes the players to two categories, one with the players with market value under 500 thousand ('under 500k') and other with players having market value above 500 thousand up to a million ('500k – 1M'). In the dataset most of the players have more than one position, so it will be easier to get only the first one displayed. Categories for each of the positions on the football pitch were made (Attacker, Midfielder, Defender, Goalkeeper). This allows analysts to focus on players they are searching for. This was done with a function that goes through all the position and checks which category each position matches. Then all the categories were implemented in a new column called 'position_category'. As the research is focused on attackers, a new data frame 'attacker_df' was created. This data frame filters the normal nacdata dataset to contain only attackers in it. To check if there are still any missing values a heatmap of them was visualized.

To see if there is any correlation between the main features of attackers, the first 20 players with the highest ‘xG’ are displayed and also their ‘Goals’, ‘Market value’ and ‘Shots on target, %’. From this table it is easy to assume that there is correlation between the values.

There are a few visuals that display a correlation between some features and distribution among them. There is a scatter plot of ‘Shots on target, %’ and ‘Goal conversion, %’, a scatter plot of ‘Height’ and ‘Weight’ and number of goals, bar plot for Goals per different Age for attackers, scatterplot of the correlation between xG and Goals. A scatter of average matches per age and same for average market value per age.

Inference from the visual and EDA is that xG is one of the most important feature for attackers and also for the research. It is correlated with actual goals, shots on target, goal conversion, and many other features that represent the performance of the attackers in the best way and can help NAC improve their attacking.

3 Machine Learning

3.1 Method

Methods used for the research are:

1. Simple Linear Regression method – This method is used for predicting the relationship between two variables using a straight line. With this model ‘xG per 90’ is predicted by the independent variable ‘Age’, which represents the Age of each player.
2. Multiple Linear Regression method – The key features in this method are more than one, compared to simple linear regression method, so the predicted variable(xG per 90) is more accurate. Using more independent variables can lead to better prediction of the dependent value.
3. Logistic Regression method – This method is using dependent variable that has only two possible outcomes(Market_Value_bin). Based on the independent variables, the dependent is predicted and goes from 0 to 1 that represent the two categories of it. The method use most of the attackers features, such as ‘xG per 90’, ‘Goals’, ‘Successful attacking actions per 90’, etc. and predict the ‘Market_Value_bin. Reg plot with ‘xG per 90’ on the x-axis and predicted ‘Market_Value_bin’ on the y-axis is visualized
4. Random Forest Classifier – RandomForestClassifier is a method that builds multiple decision trees and merges them together. This method is used because it provides a measure of feature importance and gives understanding of which of the features contribute the most to the predicted value (‘Successful attacking actions per 90’). Bar plot with feature importance is visualized.
5. Decision Tree Classifier and Regressor - The visualization of feature importance for both the DecisionTreeClassifier and DecisionTreeRegressor allows for a clear understanding of which features are influential in making predictions. The dependent variable for these methods is ‘xG per 90’ and it is plotted into Regression Scatter plot of Actual vs Predicted values, where they have strong correlation.
6. Gradient Boosting Regressor method – This learning process often leads to improved predictive performance, making Gradient Boosting a suitable choice when high accuracy is desired. However, it depends on the input data, that in the case of the research it gives very high mse. This means that the model is not that accurate.
7. XGB Classifier – This method is mostly used because of its high predictive accuracy and that can easily compute non-linear relationships and interactions between features. In the research ‘xG per 90’ is predicted from the other attackers’ stats and feature importance of XGB Classifier is plotted.

8. SVM(SVC) – SVC method is tested because in the research, the only dependent variable is 'xG per 90'. However, SVC might not be the best choice for very large datasets as in this case due to its computational complexity.
9. KMeans – In the research, KMeans is used to explore the inherent structure in the data without relying on predefined categories. It is also used to group the data into clusters and in the improvement of the model
10. Random Forest Classifier – This model is used in the Feature Selection step of the research to find the highly correlated features and correlation analysis of the model

3.2 Model evaluation

The metrics used for evaluation can be classified into Classification Metrics and Regression Metrics, depending on each model's type.

For the Classification machine learning methods, the following metrics are used:

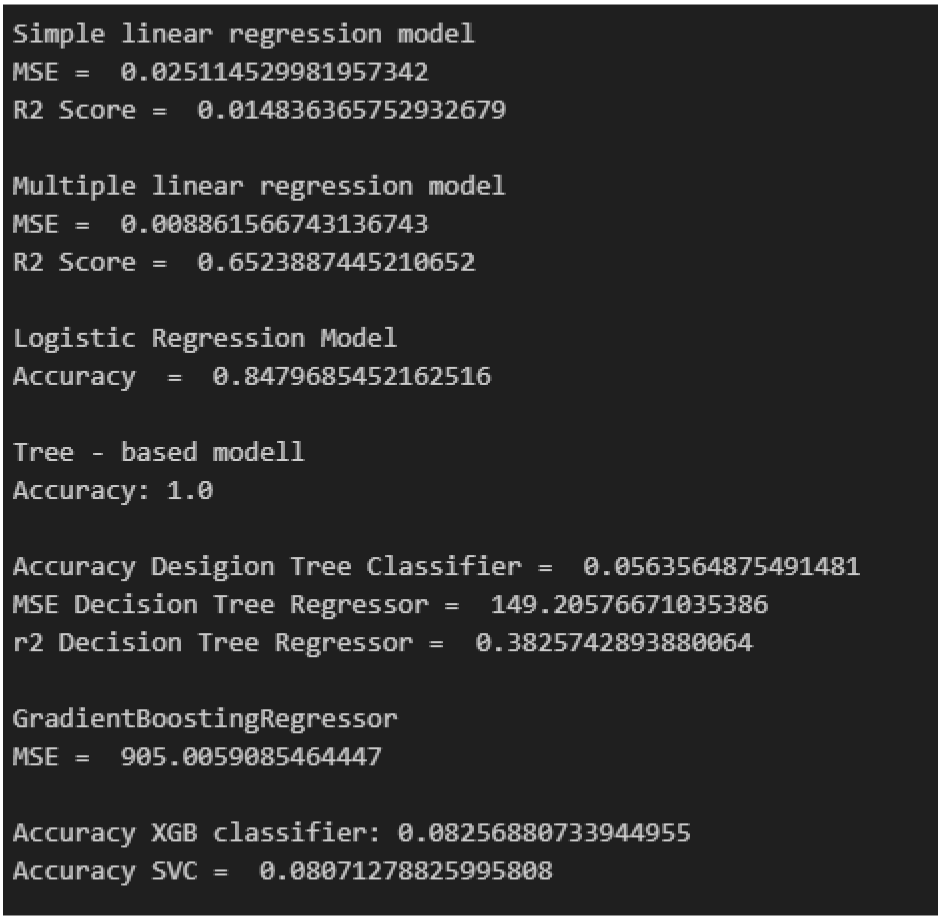
- accuracy_score – It is a fundamental classification metric that provides an overall measure of how well a model is performing. It gives information about the accuracy of the model and it is easy interpret. The closer it is to 1, the more accurate the model is.
- precision_score – It is a classification metric that measures the accuracy of the positive predictions made by a model. It is used because it is important when the goal is to minimize the number of false positives. However, precision can be high even if the overall accuracy of the model is low.
- recall_score – It is a classification metric that focuses on the ability of a model to capture all relevant instances of a positive class.
- f1_score – It is a is a classification metric that combines precision and recall into a single value, providing a balanced measure of a model's performance. Useful in situations where both precision and recall are important, and there is a need to find a balance between false positives and false negatives.

For the Regression machine learning methods, the following methods are used:

- Mean Squared Error (MSE) – It measures the average squared difference between the predicted values and the actual values in a regression task. The lower the value is, the better prediction accuracy.
- r2_score – It measures the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression task. A high R2 score implies that a significant portion of the variability in the target variable is explained by the model.

The purpose of using all the metrics is to evaluate the performance of each model.

Results from the executed machine learning models are presented here:



```
Simple linear regression model
MSE =  0.025114529981957342
R2 Score =  0.014836365752932679

Multiple linear regression model
MSE =  0.008861566743136743
R2 Score =  0.6523887445210652

Logistic Regression Model
Accuracy  =  0.8479685452162516

Tree - based modell
Accuracy: 1.0

Accuracy Desigion Tree Classifier =  0.0563564875491481
MSE Decision Tree Regressor =  149.20576671035386
r2 Decision Tree Regressor =  0.3825742893880064

GradientBoostingRegressor
MSE =  905.0059085464447

Accuracy XGB classifier: 0.08256880733944955
Accuracy SVC =  0.08071278825995808
```

Interpretation of the evaluation results:

- The R2 score of Simple Linear Regression is low, indicating that the model is not explaining much of the variance in the data.
- The multiple linear regression model has a lower MSE and a higher R2 score compared to the simple linear regression, suggesting better performance and better explanatory power.
- The logistic regression model has an accuracy of 84.93%, indicating good performance in classification.
- The decision tree classifier achieved perfect accuracy, which could indicate potential overfitting on the training data.
- The decision tree regressor has a low accuracy and a relatively high MSE, suggesting that it may not be capturing the underlying patterns well.
- The gradient boosting regressor has a high MSE, indicating that the model's predictions deviate significantly from the actual values.
- The XGBoost classifier has a low accuracy, suggesting that it may not be performing well on the given data.
- The support vector classifier has a low accuracy, similar to the XGBoost classifier.

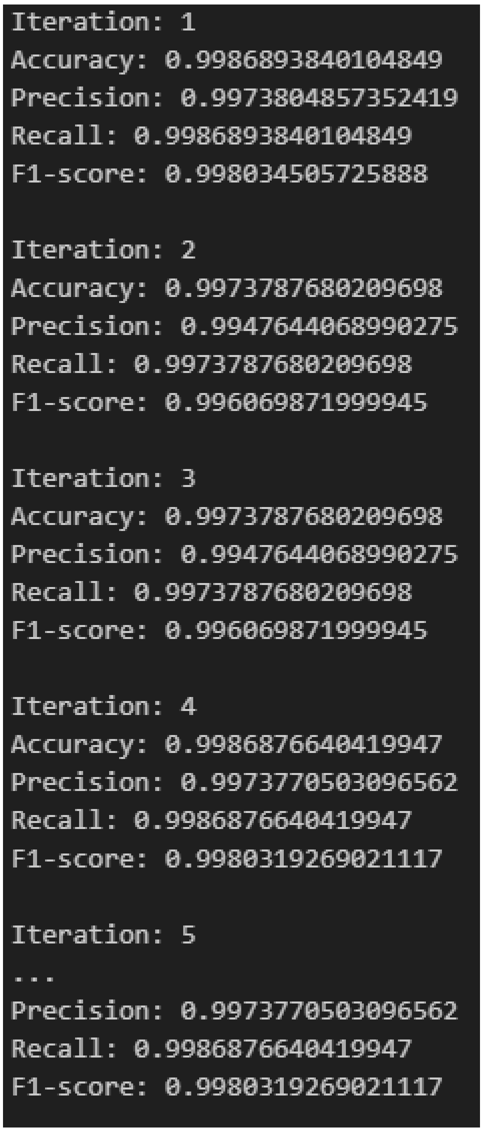
3.3 Model improvement

Key Hyperparameters focused on during tuning:

- n_estimators – The number of trees in the forest
- max_depth – The max depth of the

- min_samples – The minimum number of samples required to split an internal node

Teqnique used for hyperparameter optimization is Grid Search Cross-Validation that evaluates all combinations of hyperparameter values within the specified grid to find the optimal set. During the model improvement the DataFrame had to be with only chosen and suitable independent variables with which to achieve good results. The impact of hyperparameter adjustments on the model performance can be easily seen from the results, which lead to higher accuracy and overall performance of the methods.



```
Iteration: 1
Accuracy: 0.9986893840104849
Precision: 0.9973804857352419
Recall: 0.9986893840104849
F1-score: 0.998034505725888

Iteration: 2
Accuracy: 0.9973787680209698
Precision: 0.9947644068990275
Recall: 0.9973787680209698
F1-score: 0.996069871999945

Iteration: 3
Accuracy: 0.9973787680209698
Precision: 0.9947644068990275
Recall: 0.9973787680209698
F1-score: 0.996069871999945

Iteration: 4
Accuracy: 0.9986876640419947
Precision: 0.9973770503096562
Recall: 0.9986876640419947
F1-score: 0.9980319269021117

Iteration: 5
...
Precision: 0.9973770503096562
Recall: 0.9986876640419947
F1-score: 0.9980319269021117
```

4 Ethical Considerations

According to the official NAC Breda website the goals for the club are to achieve more success in the upcoming year (2024) and to keep the fans interested in NAC's matches. The club aims to investigate possible player transfers and find the best ones to become on the top of the league, the team is playing in. For this purpose, NAC needs a lot of personal data about the players. Apart from new players, the club is also interested in the fans views of its development and what they want to see in the future. Everything can be achieved by collecting an appropriate data needed. Using personal data in sport analytics can contain a lot of benefits for the team and its fans. Information, such as, player performance and fan engagement could be easily seen and very useful. In the website of the club, the privacy statement and collected data from NAC are shown.

As easy reachable from the website, the collected data is more likely to cover the GDPR (General Data Protection Regulation) regulations. It is widely described which data are being used and their privacy, stored from the club. The team offers a transparent and clear overview of: for what purpose the personal data is being processed, which of the personal data may be processed, how long the data will be stored, what measures the club has taken to protect everyone’s personal data. NAC also displays people’s rights of viewing, modifying or deleting their data from the site. Everyone whose information is in the site is allowed to contact NAC and describe whether some of the information or all of it should be modified or fully deleted.

Specifically, everything about collected data is being described in the privacy statement. But there are other issues that could be found with sharing the data. For example, as said in the privacy statement, the site cannot investigate if the person that data is being collected is under 16-years-old or greater. All this can lead to violation of the ethical norms and GDPR. This problem can be solved by an agreement from each visitor that has to be at least 16-years-old, so its data to be stored or not. NAC should not be allowed to use any type of personal data, without the person’s knowledge. The club has to use only the information needed for purposes of the team, its development and growth which cover the ethical principles and norms. Before collecting the information, everyone needs to be asked and assign and agreement for giving the data. When the privacy statement is being updated, all the people have to be informed and make an agreement again. The details may only be shared with individuals on the NAC staff who are authorized to access the personal data. Authorization should only be granted to individuals who have good reason to process the data. As a rule, it is not legal to publish sensitive, personally identifiable research data without the consent of the individual. NAC must strictly follow these rules.

In general, to say whether the data privacy is on top level. A lot of factors can lead to violation of the ethical norms. People cannot be sure if their privacy information is being used in appropriate way, or someone is collecting it for its own purpose. NAC may be allowed to use any type of personal data in their dashboard, only if permission is granted from the person. Any use of data without permission crosses every ethical line and do not cover the ethical norms. Situations as stealing data and using it wrong can get everyone into a lot of troubles and worries. Avoiding that it is preferable everyone to pay a lot of attention on what information gives and strictly follow who and how is likely to use the data.

5 Recommendations

Based on the findings from the data analysis and machine learning models the recommendations for NAC Breda are to focus on recruiting attackers with high expected goals (xG) as it is a significant predictor of actual goals. NAC should Continue leveraging data analytics to assess player performance, but ensure compliance with GDPR. Engage with fans to gather insights on their preferences and expectations. Consider surveys or feedback mechanisms to enhance fan experience and maintain their interest.

Strictly adhere to ethical guidelines, ensuring transparent data practices and respecting individuals' privacy. Implement age verification mechanisms to comply with GDPR regulations and build trust with fans.

Games

Leisure & Events

Tourism

Media

Data Science & AI

Hotel

Logistics

Built Environment

Facility

Mgr. Hopmansstraat 2
4817 JS Breda

P.O. Box 3917
4800 DX Breda
The Netherlands

PHONE
+31 76 533 22 03
E-MAIL
communications@buas.nl
WEBSITE
www.BUas.nl

DISCOVER YOUR WORLD

Breda
University
OF APPLIED SCIENCES