More investment in young players – brighter future for NAC

The focus of this report will be on the following query:
Do young players receive insufficient recognition, or do they simply require additional training and investment to enhance their performance?

DISCOVER YOUR WORLD

Breda University
OF APPLIED SCIENCES

More investment in young players – brighter future for NAC

Victoria Vicheva

Breda University of Applied Sciences

Applied Data Science and Artificial Intelligence

Mentors for block B:

Borislav Nachev and Rebecca Borski

Retake:

May 28, 2024

1 Introduction

Professional football is an intensely competitive realm, requiring robust strategies for sustained success. The key to ensuring the success of these strategies lies in acquiring accurate information about a specific football organization, which can be derived from insights regarding their players gained over time.

For NAC Breda, a critical question arises: "Are young players undervalued, or do they simply need more training and investment for optimal performance?". But why is it so important for NAC this question to be answered? The reason is that through answering this in this report could transform the club's trajectory. This will help decisive actions on player development, investments, and strategies are pivotal for NAC's future success and talent attraction. This includes not just established players but also the younger generation, who are the club's future.

The report aims to explore data-driven insights, analysing player metrics and historical trends. A tailored machine learning model will ascertain if young players (under the age of 25) are undervalued assets, or if strategic investments can unlock their untapped potential, shaping NAC Breda's future.

2. Exploratory Data Analysis (EDA)

2.1. High-level overview of the dataset:

The size of this dataset is considerably big, as it combines most of the NAC Breda records about player metrics. After combing and creating the unit ".csv" file from all other small ".xlsx" files which were given in the third week, a further analysis was made. From this analysis it became evident that in total there are 16 535 rows and 114 columns in total. Having this much of a data, it is vital to look more deeply into it, so for example it was necessary to see how many categorical and numerical variables are present too. In total there are 89 floats: 16 integers and 9 objects.

2. 2. Steps taken to prepare the data for analysis:

The first crucial step of the data management was identifying and handling the missing values. Consequently, this was done by various functions such as "name_of_the_data. isnull (). sum()", which indicates how many missing values are in each column, and "name_of_the_data.isnull(). sum().sum()", displaying the total number of missing values.
Even though this is a big progress into the preparation of the data, outliers also need to be considered as a "problem on the way". The reason is that with them present the model will not function optimally, as they can lead to deviations, errors, and ultimately, an ineffective model in the end. To see how they could be identified, for instance, in the preprocessing of the data (from week 3) a very useful snippet of code helped to display the number of potential outliers for the numerical variable "Age" which by summarizing the whole code are exactly 25. At this stage of preprocessing, it cannot be stated whether this is considered as poor or no, it is early because the actual analysis is not started.
Furthermore, another approach of dealing with outliers was by Filling null values with 0, so there are no null values and even when we want/need we can transform the integer into a string with the str. () function.

2.3. Summary statistics of the data:

Summarizing statistics of the data is crucial for further analysis. This is because without it we will not be able to draw in the end any conclusions or they will not be helpful at all. Descriptive statistics utilize sharp and clear summary of the leading features of the dataset.
After making it clear why this summary is so important for the analysis of the NAC dataset, now the found measures of tendency and the dispersion can be shown and interpreted. They were found by using the function "print(df.describe())" which provides a look into the highlights of a particular dataset. These "highlights" are the measures of central tendency and dispersion (***Table 1***). They can tell a lot about the variability of the dataset. For instance, there are some variables

with number of 1.40 std (in “Market value”) or a greater one as 862.64 std (in “Minutes played”). This indicates that there is a greater variability in the column “Minutes played” than in “Market value”. Logically, this tells us immediately that when the ML model is being built, this variability must be considered and it might be not the best choice to use “Minutes played” as a dependent or independent variable, or any other variable in the final ML model.

Finally, here must be also mentioned the frequency counts for categorical data in the dataset so the occurrence of some patterns can be followed and further analysed. This can be performed with just one line of a code defining a function within a data frame, the desired variable/column needs to be chosen and then defining the function “value_counts()”. The chosen variable is “Team within selected timeframe” just for example how this function works and displays the frequency counts:

	Age	**Market value**	**Matches played**	**Minutes played**
count	16527.000000	1.653500	16535.000000	16535.000000
mean	25.233860	5.0866069	20.856728	1485.484366
std	4.636223	1.400836	9.196324	862.646123

Table.1**:** Descriptive statistics of some variables from the NAC dataset

2. 4. Visual techniques, description:

Working with such a huge dataset it is vital to visualize the data and its variables/ targets properly. This is because it makes the analysis way more understandable for everyone, even people who do not have any idea of Data Science and the dataset itself.

The visualizations used in this analysis were histograms (***Fig. 1)***; scatter plots, column graphs (***Fig. 2)***, etc.

The graph which had not so many, but very interesting insights and influenced subsequent analysis steps, was the one placing the market value and the different age groups against each other, so the difference between younger and older players was visible.

The age groups which were created in this analysis were young (below 25 years) and adult players (above 25 years). The graph (**Fig. *3***) indicated that the “Market value” of adult players is not that significantly higher but still sufficiently more than the one for young players. This led to considerations for the whole analysis and about building up the final model because looking at the Fig. 11 the average percent of save rate for goalkeepers in the same age categories there was not such high difference. This meant that maybe younger players are not that less skilled than the older ones. Consequently, this led to the question of this report: Are younger players underestimated and their Market value has to be higher?

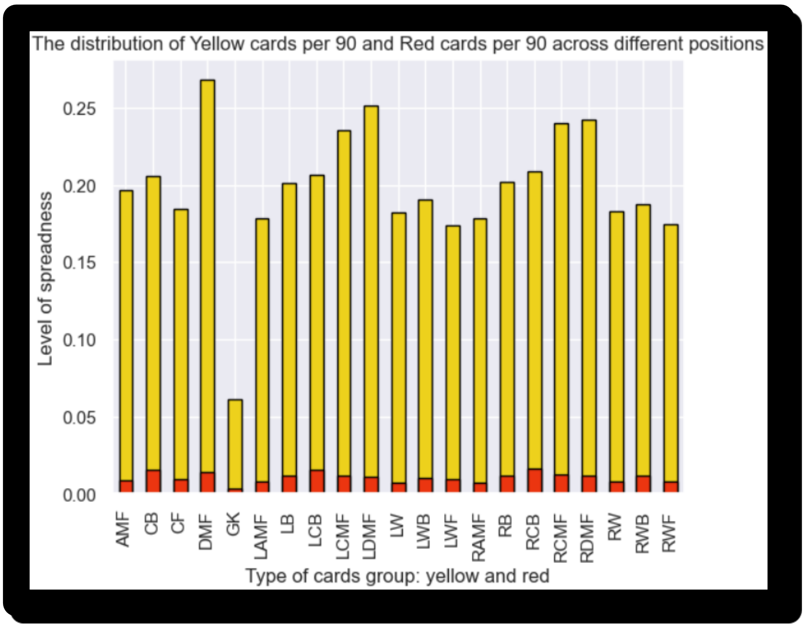


Fig.1: Example of a histogram showing the distribution of "Yellow cards per 90" and "Red cards per 90" across different positions.

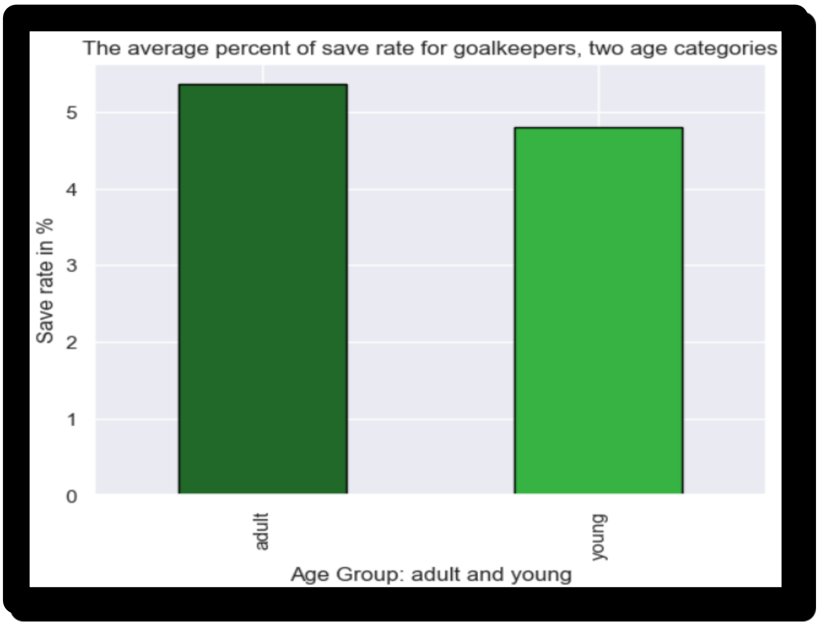


Fig. 2: Example of a column graph which indicates the trend in the "Save rate %" for goalkeepers across different age groups, young and older.

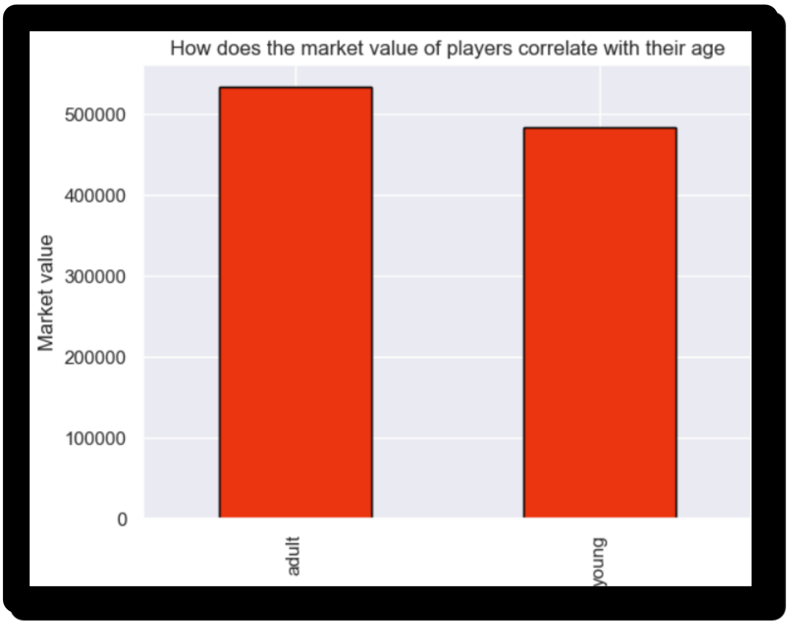


Fig.3: Comparison of the two variables “Age” (separated into two categories “young” and “adult” players) and “Market value.”

2. 5. Methods used to examine relationships between variables:

After the whole careful analysis, correlations were found and the best way of representing them was in the form of the well-known scatter plot. It can be observed that there is not a big correlation between Market value and Age. This indicates that even when the Market value is high, it is not necessary that the age goes higher too. However, what was found earlier should be kept in mind because still the older players have a higher Market value as it became clear in ***Fig. 3***, The lack of correlation does not mean there is not any, and this is why the analysis continues to observe these two variables together. The goal of this analysis is to show if younger players are underestimated or not and does NAC needs to invest more in their training or no.

2.6. Summary of key findings from the exploratory analysis:

What was found in EDA is in a way conflicting as it can be observed in ***Fig. 4*** observed the adults players have a higher Market value than the younger players and still there is not a big correlation between Market value and Age, the correlation plot indicated that when the “Age” goes higher it is not necessary that the Market values does as well.
In such huge datasets as this one, these kinds of controversial findings are normal because there is more space for deviations in the data. The best example is ***Table 2,*** where the highest Market value is the one of a 23 years old, which is above 25 and goes to the category of “young players” and the goal score is around 9. Then another example of controversial finding is again in ***Table 2*** on the 84th row where the player has a high Market value, a high Goal score of 16 and is in the category of “adult player”.

Because of this inconsistency in the data the feature selection for the final model will remain with the variables “Market value”, “Age” and “Goals”, and additionally with 'Smart passes per 90', 'Accurate smart passes, %', 'Key passes per 90', 'Passes to final third per 90','Accurate passes to final third, %', 'Passes to penalty area per 90', 'Accurate passes to penalty area, %'. Because prediction should be made on more data, so accurate conclusions can be made. Market value of a football player depends on variety of factors which are combination of performance on the field, personal characteristics and age and the general potential. (**Chen, 2024**)

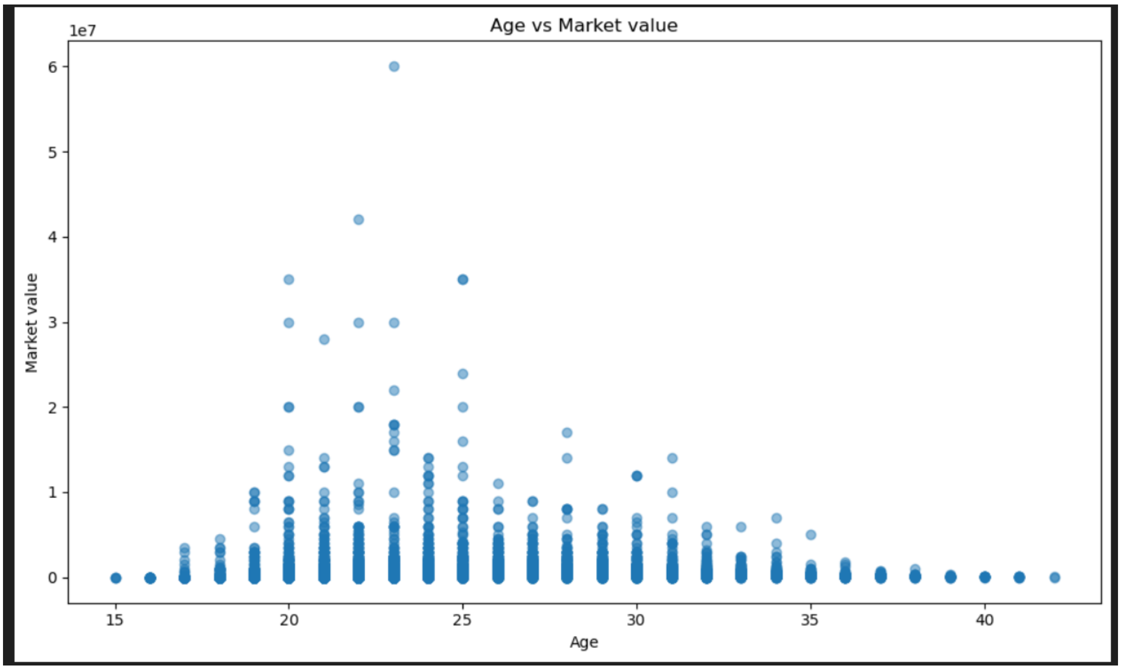
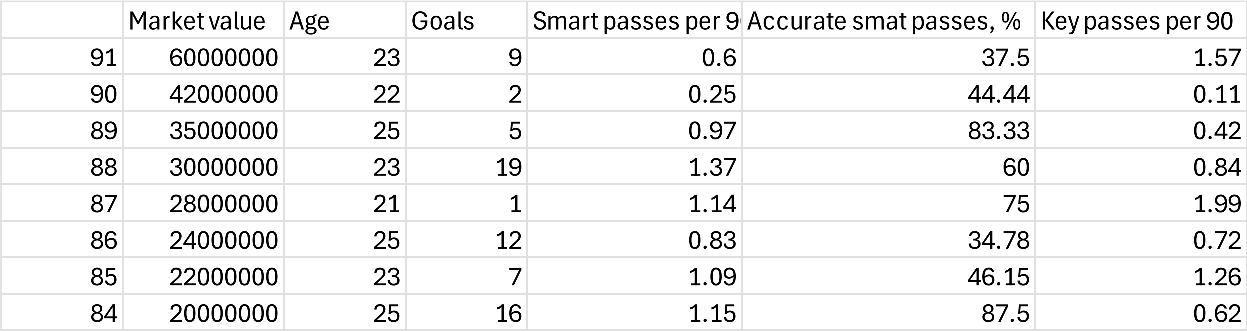


Fig. 4: *Correlation Market value Age, 0.007. This indicates weak positive correlation between the two variables.*



	Market value	Age	Goals	Smart passes per 9	Accurate smat passes, %	Key passes per 90
91	60000000	23	9	0.6	37.5	1.57
90	42000000	22	2	0.25	44.44	0.11
89	35000000	25	5	0.97	83.33	0.42
88	30000000	23	19	1.37	60	0.84
87	28000000	21	1	1.14	75	1.99
86	24000000	25	12	0.83	34.78	0.72
85	22000000	23	7	1.09	46.15	1.26
84	20000000	25	16	1.15	87.5	0.62

Table 2: *Top 8 highest Market value per player and other chosen variables*

1. Machine Learning

3.1. Chosen Model: Gradient Boosted Trees

After the detailed analysis it became clear that the variables that were chosen for this report are very contradictory to each other. Despite this, the analysis continued as it is, namely in interest to find out whether younger players are underestimated or not.
The chosen machine learning algorithm was Gradient Boosted Trees because firstly it gave the best result of around 62% and secondly because one of the advantages of decision trees is exactly that they *are capable of modeling complex and non-linear relationships between features and target variables, making them well-suited for use in a wide range of applications* **(Aksoy, 2023)**, as it is in this case.

3.2. Methods for Model Evaluation:

3.2.1. Evaluating through accuracy score:

The first method used for evaluation of the model is by printing the accuracy of the model by comparing its predictions on a test set with the actual outcomes.

Accuracy of the final model: 62%

3.2.2. Evaluating by importing confusion matrix:

This is the created confusion matrix, and it can be observed that most of the elements appear as “0,0” which indicates the number of instances that were correctly predicted as true negatives. This means that the overall performance of the model works very nice with this dataset and can predict the absence of the class of interest (**Shivaprasad, 2022**).

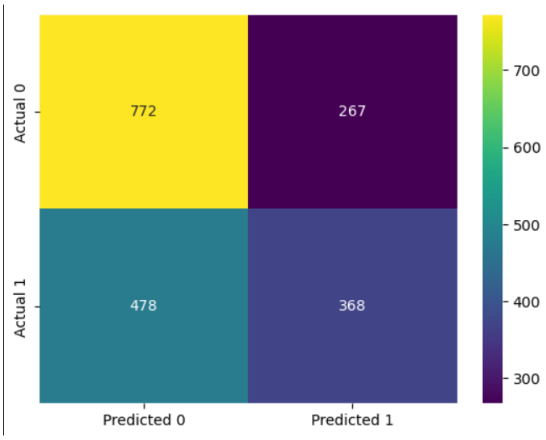


Fig.7: Confusion matrix

3.2.3. Evaluating through Precision, Recall and F1-score:

Precision:
For most of the labels it is around 0.62 (62%). This indicates that the precision of the classification model is correct, but still has room for improvement.

Recall:
Again, here for most of the labels, the recall is between 0.62 (62%). This indicates there are indeed false negatives. The positive instances in the NAC dataset have been identified correctly by the Gradient Boosting model in most cases, but still there is what to be improved.

F1-score:
One more time, also for the F1-score the labels have value between 0.62 (62%). F1-score represents how good the balance between the precision and recall is, the model here has a good performance, but as it has already mentioned ‘Age’ is a complex variable to predict and this might be the problem why this model is not so accurate.

2. Model Improvement

3.3.1. Tuning and optimization

List of hyperparameters:

‘n_estimators': [100, 150],

‘learning_rate’: [0.1, 0.2, 0.5],

'max_depth': [3, 4, 5, 6],

‘subsample’: [0.6, 0.8, 1.0],

'min_samples_split': [2, 5, 10],

'min_samples_leaf': [1, 2, 4]

Techniques used for hyperparameter optimization:
Random Search where the overall found score is sufficiently nice and efficient but the problem encountered during the process of running the code is that it took too much time and maybe if we use even larger dataset or put more data in the NAC data it will not be much of favor for us to use it.

3. Ethical Considerations

This study, which involves developing a machine learning model to analyse player data from NAC Breda, operates within a robust ethical framework. The ethical principles guiding this project are outlined below:

4.1. Ethical Company

NAC Breda's ethical standards are upheld through clear policies, including compliance with the General Data Protection Regulation (GDPR) (**Wolford, 2023**). Sensitive player data remains confidential, with no information shared beyond the scope of analysis. Transparency is maintained throughout the project, with detailed explanations provided at each stage.

4.2. Ethical process and tools

The main tools that were a part of ethical preprocessing of NAC's data are transparency and explainability.
Transparency is a core principle in the development of the machine learning model. Every step, from data preprocessing to model evaluation, is meticulously documented to ensure a clear and transparent understanding of the decisions made during the whole preprocessing of the data. Combined with explainability why certain decisions were taken to get this level of accuracy, this approach not only builds trust among stakeholders but also empowers users to assess the model's reliability and fairness.

4.3.Ethical People

The researcher (Me, Victoria Vicheva) involved in this project adhere to strict ethical standards, using player data solely for the intended analysis. The model itself integrates societal values and ethical considerations into its reasoning, ensuring responsible decision-making.

In conclusion, NAC Breda's commitment to ethical guidelines is evident throughout this study. However, as a significant organization handling sensitive data such as the main variable used in this research "Age", continued vigilance is necessary **(*Personal Data - General Data Protection Regulation (GDPR)*, 2021).** Individual consent for the processing of personal data remains a paramount consideration, ensuring that ethical standards are upheld. This means the rights of individuals are respected by all members of NAC, including people from management, players, etc., meaning all personal information is ethically used and preserved (**Connor, 2023)**

4. Recommendations

In every project of this nature, adjustments and suggestions for improved future performance are crucial. As previously mentioned, predicting variables like ‘Age’ can be challenging. This is why this variable has been categorized into ‘young players’ (under 26 years) and ‘adult players’ (over 26 years). The goal of this report was to assess whether younger players are undervalued, and indeed, they might be.

However, it is important to recognize that additional player characteristics play a significant role in determining Market value. Football teams consist of players with diverse roles and abilities, each contributing differently to their Market value.

Considering player age, it is reasonable to expect that older players generally have higher Market value, because as it was found out in the ML research, they most of them have played more minutes on the field. Yet, younger players possess greater fitness levels and tend to learn quickly. This helps explain why the player with the highest Market value is typically under 25, falling within the "young player" category.

For example, defensive players require speed and coordination more than other positions. Here is a suggestion: NAC Breda might benefit from investing in younger players where there is a need for added "power" within the team. Younger players could bring vitality and vigour to the game, while older players contribute wisdom and experience, creating a balanced and successful team dynamic.

Sources:

Aksoy, G. (2023, February 18). A comprehensive guide to Tree-Based Machine Learning Algorithms. *Medium*. https://medium.com/@grkemaksoy/a-comprehensive-guide-to-tree-based-machine-learning-algorithms-45faf662fb79

Chen, J. (2024, March 16). *What is market value, and why does it matter to investors?* Investopedia. https://www.investopedia.com/terms/m/marketvalue.asp

***Classification: True vs. False and Positive vs. Negative*. (n.d.).** Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

Connor, H. (2023, May 16). *What is sensitive information?* Hayes Connor. https://www.hayesconnor.co.uk/news-resources/news/what-is-sensitive-information/#:~:text=Under%20Article%206%2C%20processing%20of%20sensitive%20information%20is,to%20which%20the%20controller%20is%20subject.%20More%20items

OpenAI. (2024). *ChatGPT interaction on 14 May 2024 [Online forum comment]. Retrieved from* ChatGPT

***Personal Data - General Data Protection Regulation (GDPR)*. (2021, 22 oktober).** General Data Protection Regulation (GDPR). https://gdpr-info.eu/issues/personal-data/

Shivaprasad, P. (2022, March 30). Understanding Confusion Matrix, Precision-Recall, and F1-Score. *Medium*. https://towardsdatascience.com/understanding-confusion-matrix-precision-recall-and-f1-score-8061c9270011

Wolford, B. (2023, September 14). *Art. 6 GDPR – Lawfulness of processing*. GDPR.eu. https://gdpr.eu/article-6-how-to-process-personal-data-legally/

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