# Prediction of goals

Artjoms Musaelans 234535

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

# Index

**Final report “Prediction of goals”**

Artjoms Musaelans
Applied Data Science and Artificial Intelligence, Breda University of Applied Science

ENGL 999: Project 1B

MSc Bram Heijligers ; MA Zhanna Kozlova

January 16, 2024

# 1. Introduction

This study examines a football player dataset from NAC to optimize player-related decisions and enhance success in professional football.

## 1.1 Background and business problem

NAC must make strategic decisions impacting team success. This study aims to understand factors affecting a player's ability to score goals, a key indicator of success.

## 1.2 Objectives and focus

The dataset is analysed using data-driven methods to identify trends and patterns, focusing on player characteristics, goal scoring, and team interaction, providing valuable insights for team formation, player involvement, and performance improvement decisions.

## 1.3 Rationale for the chosen focus

Understanding goal-scoring dynamics is crucial for developing successful team plans, as goals impact game results and player efficacy, thus emphasizing their importance to the club's main goal of winning football games.

This report details dataset management procedures and machine learning techniques for forecasting, aiming to enhance NAC's success and competitiveness in the ever-changing professional football scene.

# 2. Exploratory Data Analysis

## 2.1 Overview of the dataset

The dataset analyzed includes 16535 football players, sourced from Breda University of Applied Sciences and Brightspace. The NAC club provided data for the university. The mixed dataset includes category and numerical variables, providing a comprehensive profile of each player.

## 2.2 Steps taken to prepare the data for analysis

### 2.2.1. Handling missing values

Identification and handling of missing values were performed diligently. All missing or indeterminate data for numeric values (NaN) were replaced with the median, and all missing or indeterminate data for objects were deleted because this data could not be predicted and replaced, as a result of which the dataset was devoid of missing values, which paved the way for subsequent analysis.

### 2.2.2. Handling outliers

Using statistical methods, an extensive examination was carried out to identify and handle outliers. The goal of handling outliers was to preserve the integrity of the dataset by preventing unwarranted effect on ensuing analysis.

### 2.2.3. Data transformation techniques

The dataset was transformed using techniques like normalization, standardization, and categorical variable encoding to ensure consistent numerical characteristics, ensuring compatibility with machine learning models and stable analysis for future research.

## 2.3 Summary statistics of the data

### 2.3.1. Measures of central tendency

Additional measures of central tendency, such as the mean and median, were computed for significant numerical aspects in order to provide a more thorough insight.

### 2.3.2. Mean and median market value

Mean Market Value: USD 17,536,206
Median Market Value: USD 7,500,000

### 2.3.3. Mean and median age

Mean Age: 25.57 years.
Median Age: 25.0 years.

### 2.3.4. Frequency counts

The dataset's distribution of player positions should be highlighted to highlight the predominant positions. Individual players are classified based on their positions, providing a detailed analysis of the NAC's data, allowing a sophisticated understanding of the distribution and frequency of player positions.

## 2.4 Visual techniques used to understand the data

A variety of visualizations, such as scatter plots, bar plots, pie charts, and box plots, were used in the exploratory investigation. These visualizations played a crucial role in unraveling insights from the dataset.

### 2.4.1. Influence on subsequent analysis

The knowledge obtained from the visualizations had a major impact on later studies. For instance, a more thorough investigation into the dynamics of player valuation based on age was spurred by the correlation scatter plot between market value and age, which resulted in more complex patterns and a better comprehension.

## 2.5 Methods Used to Examine Relationships Between Variables

### 2.5.1. Correlation coefficients

A correlation analysis revealed a statistically significant association between market value and age, but the correlation coefficient is not strong (-0.06), suggesting a need for a more comprehensive analysis of variables affecting player valuation.

### 2.5.2. Examination of significant correlations

It is important to pay attention to the crucial correlation between height and weight. The strong positive correlation (0.79) suggests a discernible trend – as player height increases, their weight tends to increase proportionally. This insight could be crucial in understanding the physical attributes of players.

## 2.6 Summary of key findings from exploratory analysis

### 2.6.1. Potential hypotheses

Potential theories begin to take shape based on the observations gathered. For example, a deeper examination of the strategic preferences of top teams may be necessary given the prevalence of particular positions on these teams. Furthermore, the relationship between player performance measurements and physical characteristics (weight and height) may give rise to theories on how body composition affects player efficacy.

### 2.6.2. Impact on further analysis

Advanced analyses have a strong basis thanks to the results of the exploratory analysis. Feature selection, model construction, and the development of prediction hypotheses may be influenced by additional research into the detected trends, patterns, and correlations. Predictive analytics and more complex models are made possible by the connection between player traits and performance data.

To sum up, a solid foundation for further research and in-depth knowledge has been established by the data management, comprehension, cleaning, exploratory analysis, and visualizations. There is a lot of information available in this dataset, and deeper investigation should reveal more complex features and trends.

# 3. Machine Learning

## 3.1 Method

Linear Regression was the chosen machine learning approach for the goal of creating a prediction model. The rationale behind selecting this approach is its ability to model linear connections in the dataset, which is very useful when forecasting a numerical result like player goals. The effectiveness of linear regression in managing the linear relationships between features and the target variable is well-established, and it is also quite simple and easy to understand.

## 3.2 Model Evaluation

The linear regression model's effectiveness was evaluated using measures like Mean Squared Error (MSE) and R-squared, which are relevant to numerical value prediction. R-squared measures the model's data explanation, while MSE shows the average squared discrepancies between expected and actual values. Due to the dataset's size and resource constraints, cross-validation procedures were not employed.

The assessment results indicate that the Linear Regression model performed admirably, producing a high R-squared value of 0.9902806533224722and a low MSE of 0.06909808839196138. The combined value of these measures indicates how well the model captures the volatility in goal counts and produces precise forecasts.

## 3.3 Model Improvement

Careful examination of important hyperparameters was necessary to optimize and fine-tune the Linear Regression model. Important variables were the multicollinearity handling strategy and the regularization strength. Grid search and other approaches were used to systematically evaluate several setups in order to determine the most efficient arrangement for hyperparameter tuning.

The main challenges that arose throughout the tuning process were finding the ideal ratio between model complexity and preventing overfitting. Regularization parameters, which affect how each feature affects the model, were carefully considered in order to solve this. The hyperparameter settings almost did not change the performance of the model in any way, as evidenced by a mean square error of 0.06913776684662926 and a R-squared value of 0.9902750721455464.

This painstaking process of fine-tuning highlights how important hyperparameter selection is to be improving the predictive power of the model. The best machine learning results may be obtained by making systematic parameter tweaks, even though, in this case, the adjustments did not lead to significant improvements in the Linear Regression model.

# 4. Ethical Considerations

This project, which involves developing a machine learning model to evaluate player data from NAC Breda, is carried out under a stringent ethical framework. The ethical reasons that underpin this endeavour are outlined below.

## 4.1 Ethical company

NAC Breda's ethical standards are supported by defined policies and procedures that assure adherence to ethical norms and legislation, such as the General Data Protection Regulation (GDPR). The company promotes a culture of transparency, inclusivity, and continual feedback.

GDPR compliance: NAC Breda guarantees that all data handling activities adhere to GDPR, which requires rigorous data protection and privacy protections for EU citizens (European Parliament and Council of the European Union, 2016). This includes:

- Data minimization: Only the data required for the project is collected and used.
- Consent: Informed consent is obtained from players before data collection. They are informed about the purpose, methods, and potential uses of their data.
- Anonymization: Personal identifiers are removed or masked to protect player identities.

Transparency and feedback: NAC Breda keeps open lines of communication with staff, players, and external stakeholders to ensure openness and continual progress.

- Regular updates: Provide regular updates on the project's progress and any changes to data management practices.
- Stakeholder meetings: Hosting meetings with stakeholders to discuss the project and solicit input.
- Feedback systems: Setting up systems for stakeholders to submit feedback, which is then reviewed and incorporated into the project procedures.

Responsible parties: The data management and analytics team are responsible for implementing and maintaining these ethical standards. This team includes data scientists, legal advisors, and privacy officers.

## 4.2 Ethical Process & Tools

The ethical treatment of NAC data emphasises transparency, explainability, and the incorporation of ethics into development processes.

Transparent processes: NAC Breda provides clear and accessible information on its data gathering, analysis, and usage methods. This includes:

- Documentation: Detailed documentation of data sources, methods, and transformations used on the data.
- Communication: It is important to communicate these processes to stakeholders on a regular basis so that they understand how their data is used.
- Public reports: Publishing reports outlining the project's techniques and ethical considerations.

Explainability of judgements: NAC Breda makes certain that the reasoning behind data-driven judgements is readily conveyed. This involves:

- Algorithm transparency: Explaining the algorithms utilised, including their design, usefulness, and limits (Doshi-Velez & Kim, 2017).
- Decision rationale: Providing concise explanations for data-driven decisions, highlighting the factors examined and their impact on outcomes.
- Impact assessment: Evaluate and communicate the potential effects of these decisions on players and the organisation.

Ethical concerns are incorporated at all stages of the development process to ensure conformity with the organization's ethical standards and social values (Floridi et al., 2018). This includes:

- Ethical evaluations: Conducting regular ethical evaluations of data management procedures.
- Ethical frameworks: Using existing ethical frameworks to guide the development process.
- Ethical training: All team members will receive training on ethical development techniques.

Proof of ethical processes and tools:
To ensure ethical decision-making within this project, I followed the framework outlined by Dignum (2019) in chapters 3 ("Ethical Decision-Making") and 4 ("Taking Responsibility"). Here are the steps and tools implemented:

- Chapter 3: The decision-making process involved identifying ethical challenges, assessing stakeholder impacts, and evaluating alternatives. Risk evaluations, stakeholder mapping, and engagement meetings addressed data privacy, bias, and openness concerns. An ethical review board ensured ethical norms were met and processes were documented for accountability.
- Chapter 4: The ethical monitoring committee, consisting of Data Scientists, DPO, Legal Advisor, Chair, Stakeholder Engagement Officer, and Ethical Trainer, ensured ethical data analysis, GDPR compliance, legal guidance, evaluations, stakeholder communication, and ethics training. Regular audits and feedback loops improved processes, and guidelines based on Dignum's framework were distributed. Monitoring systems ensured compliance and corrective actions in case of breaches, promoting accountability and transparency.

Responsible parties: NAC Breda's professionals oversee the project's ethical processes, with the data protection officer (DPO) ensuring GDPR compliance, the data scientist ensuring transparency, the ethics committee conducting regular evaluations, the legal advisor ensuring legal conformity, the data privacy analyst securing data, and the ethical trainer providing ongoing training.

## 4.3 Ethical people (Employees and clients)

Professionals and clients must behave ethically towards their stakeholders. This covers their interactions with consumers, owners, society, the environment, and suppliers, as well as their understanding of ethics and moral responsibility.
Ethical professional conduct: As the researcher, me, Artjoms Musaleans follows strict ethical rules. This includes:

- Confidentiality: Using player data just for the intended analysis and keeping it confidential.
- Integrity: Ensuring that all data processing and analysis procedures are carried out with integrity and honesty.
- Ethical decision-making: Making judgements based on society norms and ethical considerations.

Awareness and Training: All team members receive ongoing training on ethical standards to ensure they are prepared to handle sensitive data ethically and make educated judgements. This involves:

- Ethical standards: GDPR training, ethical data management standards, and responsible decision-making.
- Ongoing education: Offering materials and training on ethics in data science and machine learning.

Responsible Decision-Making: Decision-making procedures include ethical concerns, which balance the interests of all stakeholders. This includes:

- Stakeholder analysis: Evaluating the impact of data consumption on players while protecting their rights and privacy.
- Ethical committees: Forming committees to examine and steer decision-making procedures.
- Impact assessments: Conducting impact analyses to better comprehend the broader ramifications of data-driven decisions.

Responsible parties: All team members, with oversight from the ethical committees, are responsible for adhering to these ethical standards and ensuring responsible decision-making.

This research highlights NAC Breda's commitment to ethical ideals, emphasizing the need for constant monitoring, particularly for sensitive data processing, individual consent, respect for personal rights, and continuous progress in ethical procedures and stakeholder involvement.

**Recommendations to Improve Ethical Standards:**

- Regular ethics training should be provided to all data administration and analysis workers, covering GDPR compliance, data protection best practices, and ethical decision-making frameworks (Floridi et al., 2018).

- Ethical audits are crucial in ensuring adherence to ethical standards and identifying opportunities for improvement. They can detect unintentional deviations from norms and offer actionable recommendations for enhancing ethical behaviour (European Parliament and Council, 2016).

- Engaging stakeholders in discussions on data ethics fosters a culture of continuous improvement, involving feedback from gamers, staff, and other stakeholders in decision-making processes on data management procedures.

- An ethical oversight committee, consisting of data scientists, legal specialists, and ethicists, should be formed to oversee the project's ethical issues, advise on ethical problems, examine data management policies, and ensure the research meets the highest ethical standards. (Floridi et al., 2018).

# 5. Recommendations

Drawing insights from the extensive analysis conducted to address the problem statement outlined, several key recommendations emerge for the client, the NAC.

## 5.1 Data management and understanding

### 5.1.1. Enhance data documentation

The NAC should implement a more robust system for documenting the data collection process, including detailed information on sources, methods, and any transformations applied. This will contribute to greater transparency and reproducibility in future analyses.

### 5.1.2. Implement data quality assurance measures

It is essential to establish frequent validation and quality checks for data. To preserve the integrity of the dataset used for decision-making, this entails correcting missing values, outliers, and inconsistencies.

## 5.2 Exploratory data analysis and visualizations

### 5.2.1. Expand visual representation:

Even if the exploratory data analysis yielded insightful findings, stakeholders may find it easier to access information by utilizing a wider range of visualizations, such interactive dashboards.

### 5.2.2. Further investigate correlations:

Nuanced patterns can be uncovered by a more thorough investigation of the interactions between the variables, particularly using sophisticated statistical techniques. This has the potential to enhance comprehension of the variables impacting the intended results.

## 5.3 Machine learning

### 5.3.1. Explore alternative models:

Think about exploring different machine learning models. Certain approaches, such as deep learning or ensemble techniques, can offer new insights and even improve prediction accuracy.

### 5.3.2. Continued model evaluation:

Establish a method for ongoing model review that includes evaluating the model's performance on a regular basis using new data. This will guarantee the long-term relevance and dependability of the model.

## 5.4 Ethical considerations

### 5.4.1. Strengthen conflict of interest policies

Address instances of possible conflicts of interest that have been found by strengthening and improving organisational policies. This includes laying out in more detail the kind of professional ties that are acceptable within the club.

### 5.4.2. Integrate ethical training programs

Create and put into place continuing ethical education initiatives for employees at all levels. The ethical standards, frameworks for making decisions, adherence to GDPR regulations, and recommendations for statistical practice should all be reinforced by these programmes.

## 5.5 Overall organizational recommendations

### 5.5.1. Establish a cross-functional data team

Assemble a committed, interdisciplinary group of data scientists, subject matter experts, and legal and ethical experts. This group can work together to solve problems using data, think through moral dilemmas, and streamline the decision-making process.

### 5.5.2. Encourage a culture of continuous improvement

Foster a culture within NAC that values continuous improvement. Encouraging feedback loops, consistent training, and the application of best practices will guarantee that data procedures and ethical standards are continuously improved.

The combined goal of these suggestions is to steer the NAC club in the direction of a decision-making structure that is more operationally effective, morally sound, and data driven. The club's goal of greatness on and off the pitch will be furthered by putting these proposals into practice.

# Literature

American Statistical Association. (2018). Ethical guidelines for statistical practice. Retrieved from https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/pdf/1702.08608

Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer.

European Parliament and Council of the European Union. (2016). General Data Protection Regulation (GDPR). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.

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