Revolutionizing Player Recruitment: A Data-Driven Approach for NAC

Leveraging Machine Learning and Advanced Analytics to Transform Talent Identification and Acquisition

DISCOVER YOUR WORLD

Data Science & Artificial Intelligence

Anastasiia Mokhonko

Breda University OF APPLIED SCIENCES

Index

1 Introduction

1.1.1. Introduction

In football, player recruitment plays a significant role in a team's success. However, traditional methods of relying on scouts' subjective judgments can lack precision. This is where data-driven analysis comes in.

1.1.2. Problem statement

NAC, like many other teams, has the challenge of identifying players who can contribute significantly to their goals. The core issue is the optimization of player recruitment. NAC needs to assess players accurately, particularly in goal-scoring, which is a crucial factor in football success.

1.1.3. Solution Overview: Predictive Modelling for Goal Scoring

Developing a predictive model using Gradient Boosting, a statistical technique to aid NAC in their recruitment process. The model analyses player data to predict their ability to score goals.

1.1.4. Objective

The primary goal is to improve NAC's player recruitment. The model offers accurate predictions on players' goal-scoring potential, helping make informed recruitment decisions.

1.1.5. Model's Role in Recruitment and Strategy

The model not only evaluates players but also helps in planning team strategies. It helps NAC understand how a player could contribute to the team's goal-scoring, influencing recruitment and tactical decisions.

1.1.6. Report Outline

The report explains the model's development process, including data analysis, the Gradient Boosting technique details, how the model was evaluated, and ethical considerations. It concludes with recommendations for implementing the model in NAC's recruitment and strategy planning.

2 Exploratory Data Analysis

1.1.7. Overview of the dataset and the steps taken to prepare the data for analysis:

A large dataset consisting of 16,535 entries spread across 114 columns sourced from 45 unique files was used to build a football analytics model. The dataset contained a vast amount of football-specific data, with a majority of numerical columns (105) and a few categorical ones (9), which was crucial for accurately predicting goals.

A thorough data cleansing process was followed to ensure the dataset's quality and reliability. The process involved several crucial steps, such as systematically labelling missing entries in categorical data as 'Unknown' to maintain consistency across the dataset and filling in missing values with zeros for numerical data to ensure uniformity.

Duplicate entries were identified and removed to preserve the integrity of the data. This step was crucial to prevent any skewed analysis results. Inconsistencies, such as extra spaces in categorical columns, were removed to avoid misinterpretation of the data.

Particular attention was paid to columns with significant missing values, specifically those with exactly 232, 1510, or over 15255 missing entries. Rows that had missing values in these critical columns were removed. Key columns like 'Contract expires', 'Team', and 'Foot' had missing entries filled in with default values like 'Not Specified', 'No Team', and 'Not Known'. Zero was used as a placeholder for columns with a very high number of missing values.

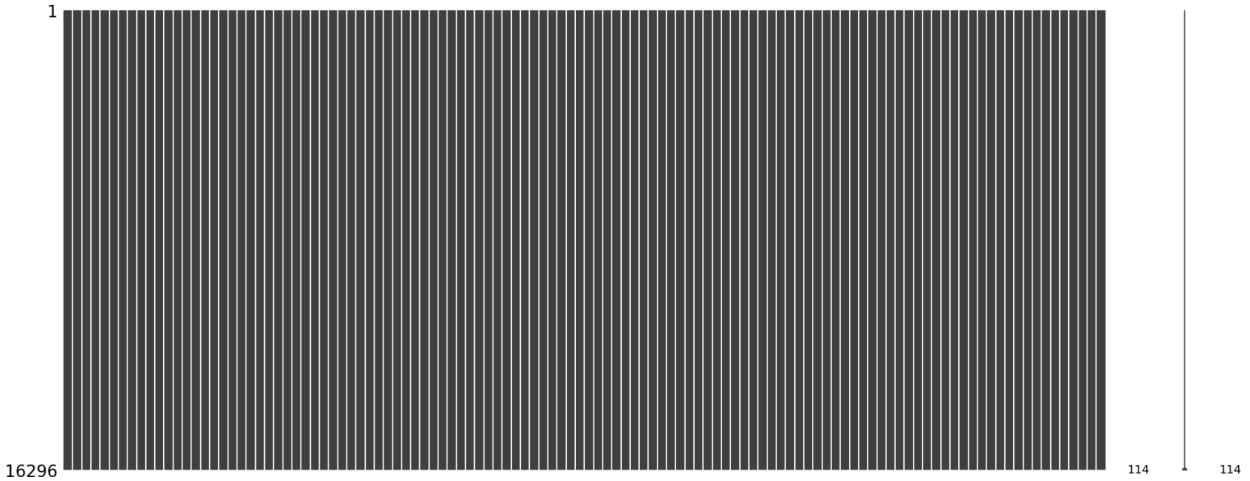
This comprehensive approach to data cleansing was vital in preparing a complete and consistent dataset, setting a solid foundation for subsequent analysis and modelling to predict goal-scoring outcomes accurately.

1.1.8. Summary statistics of the data, visual techniques used to understand the data and methods used to examine relationships between variables:

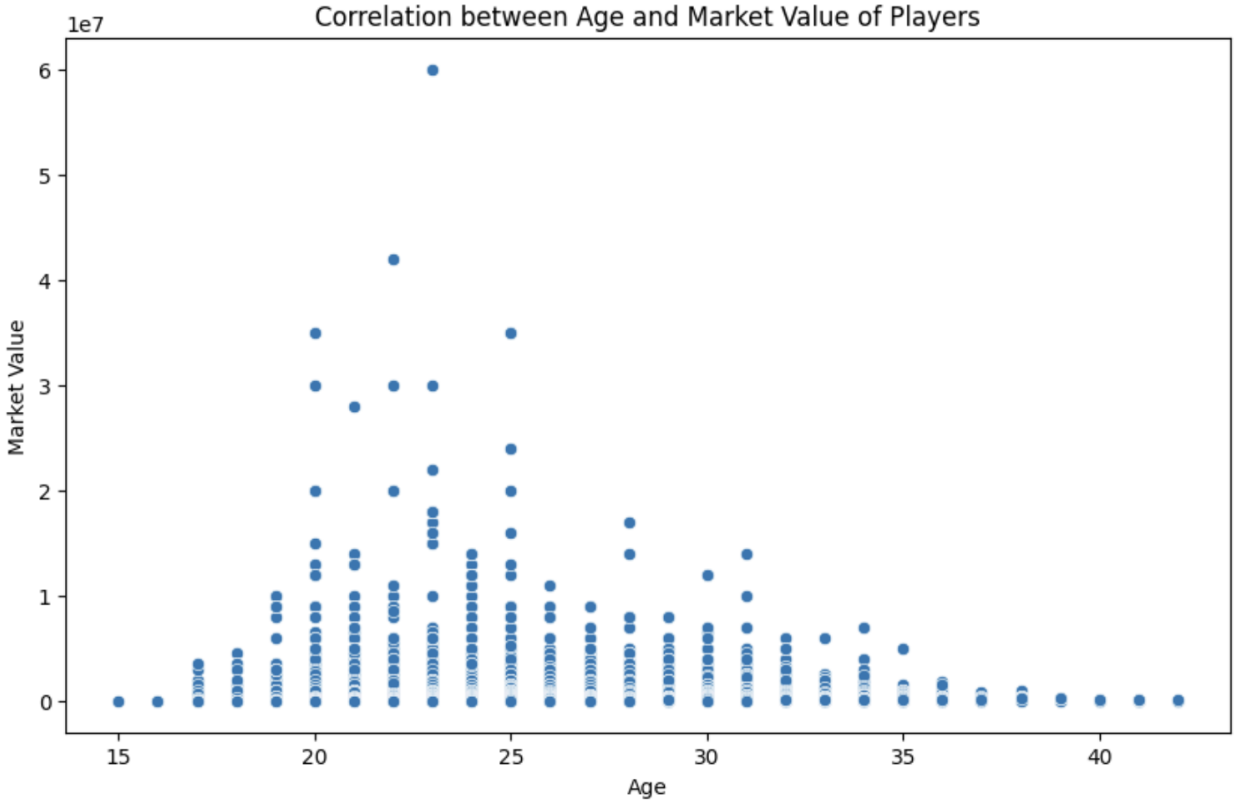
Various visual techniques were used to understand the dataset comprehensively. Each technique offered unique insights. For instance, histograms allowed the analysis of the distribution of players' ages and market values, which was crucial in identifying key patterns and outliers. Scatter plots were instrumental in exploring relationships between different player attributes, providing a nuanced view of player characteristics. Box plots enabled the examination of the range and distribution of player statistics, such as goals scored and assists, and were particularly useful for spotting outliers and understanding the overall spread of the data.

A thorough correlation analysis was conducted to calculate correlation coefficients for numerical features. This analysis revealed significant correlations between offensive metrics like goals and assists, highlighting how different aspects of a player's performance interrelate and their combined impact on goal-scoring potential. The correlation between passing behaviour and goal-scoring opportunities was also a fundamental discovery, providing insights into the multifaceted nature of player performance.

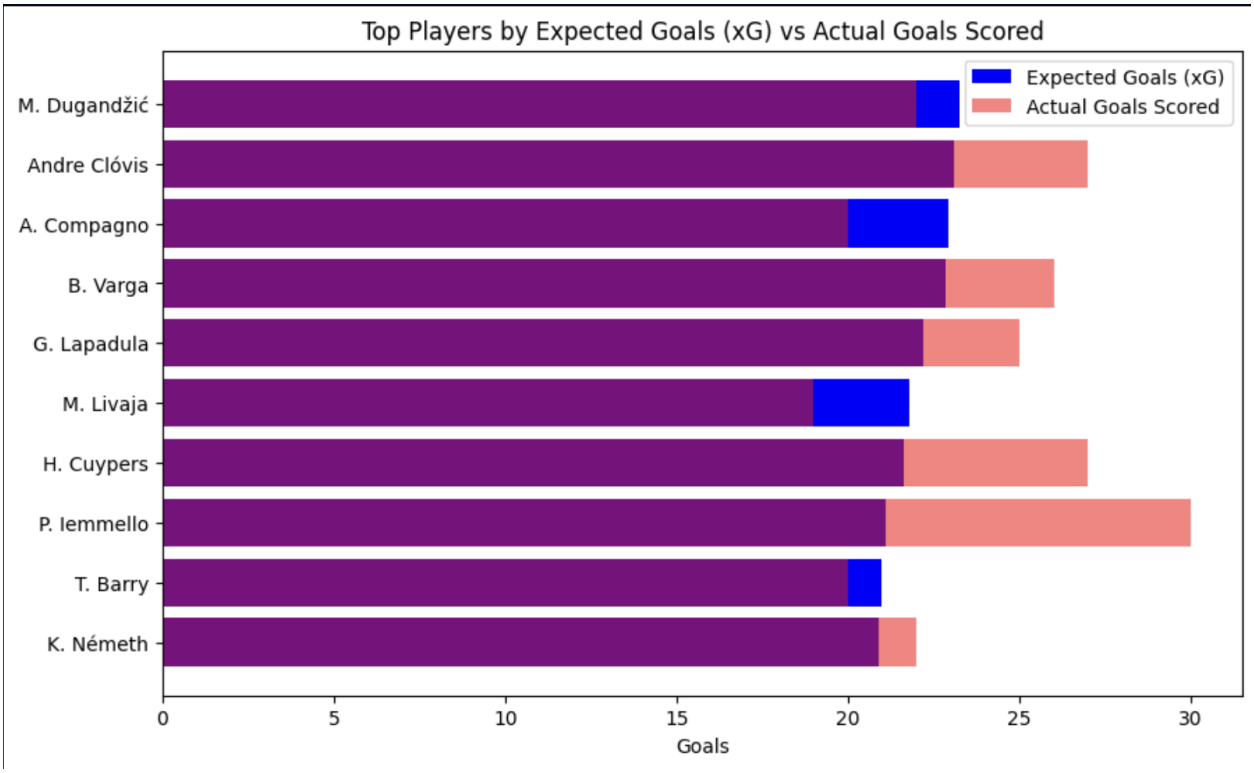
The key findings from this exploratory analysis formed the foundation for subsequent steps. Significant trends and patterns, such as the correlation between offensive metrics, informed hypotheses regarding goal-scoring abilities, which directly influenced further data analysis, particularly in feature selection for predictive models. By identifying significant correlations and patterns, it was possible to more accurately determine which features were most relevant to predicting goal-scoring outcomes. This approach ensured that the model was built on a robust and insightful understanding of the dataset, laying the groundwork for accurate and effective predict ions.



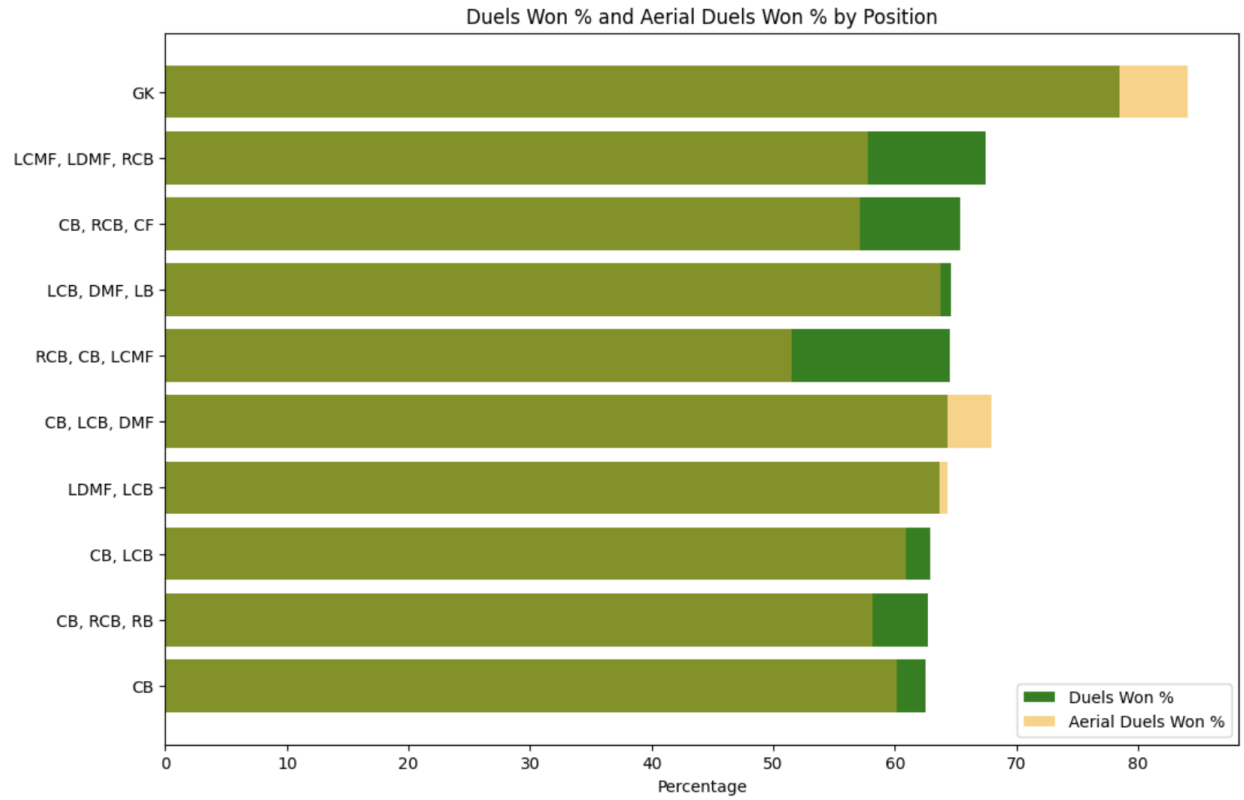
(The result of cleaning the dataset)



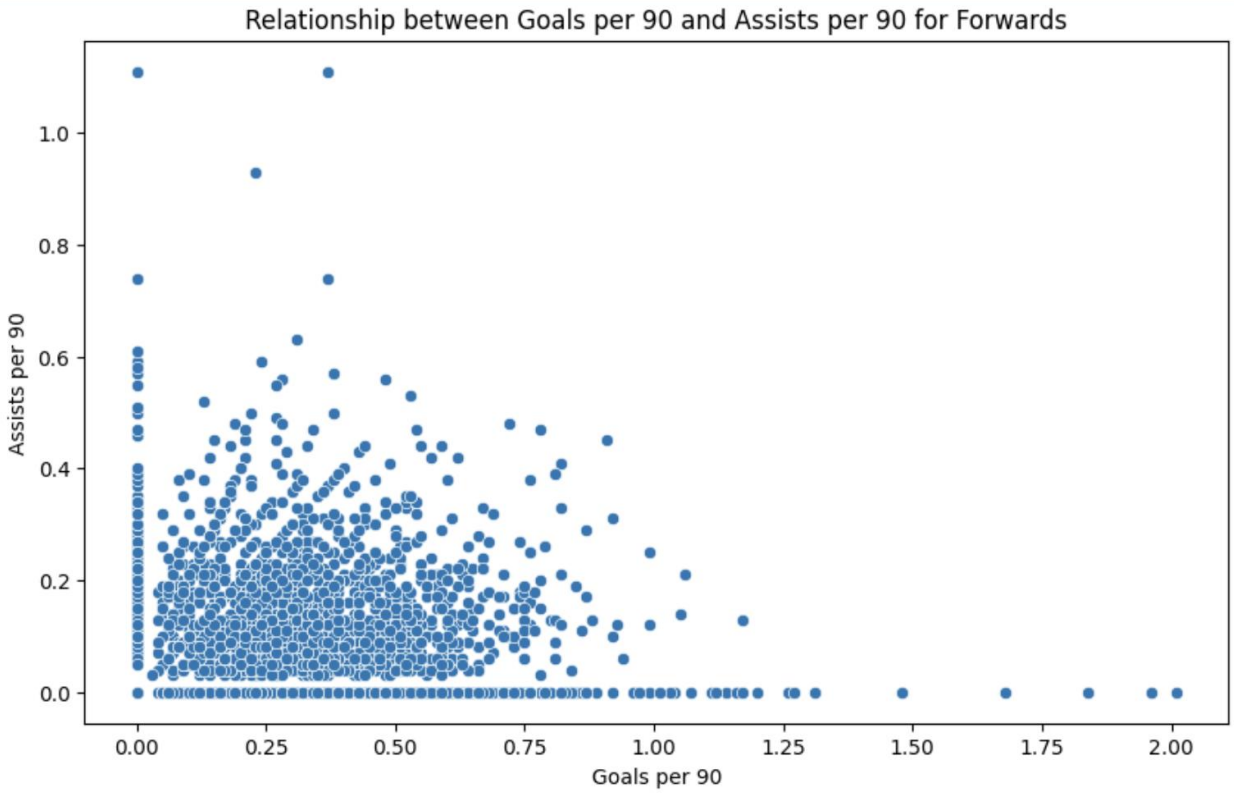
(Correlation between Age and Market Value of Players)



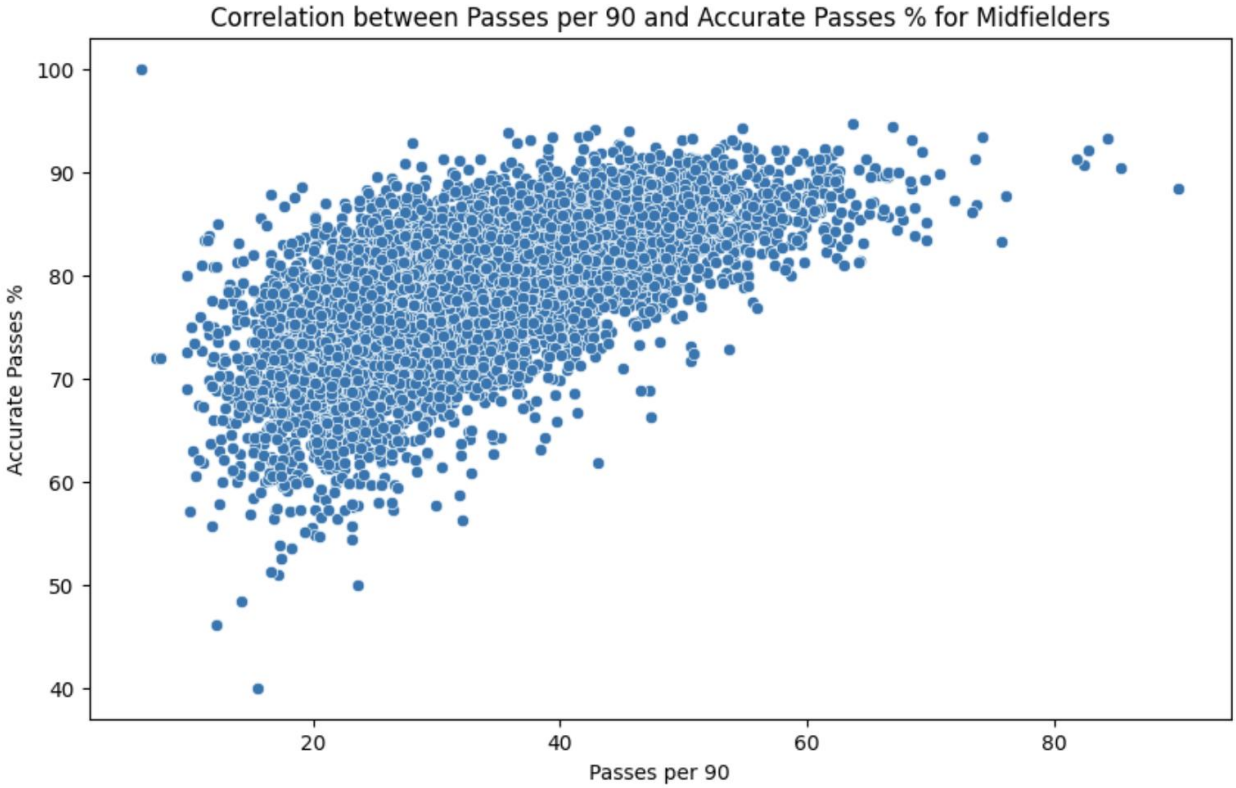
(Top Players by Expected Goals(xG) vs Actual Goals Scored)



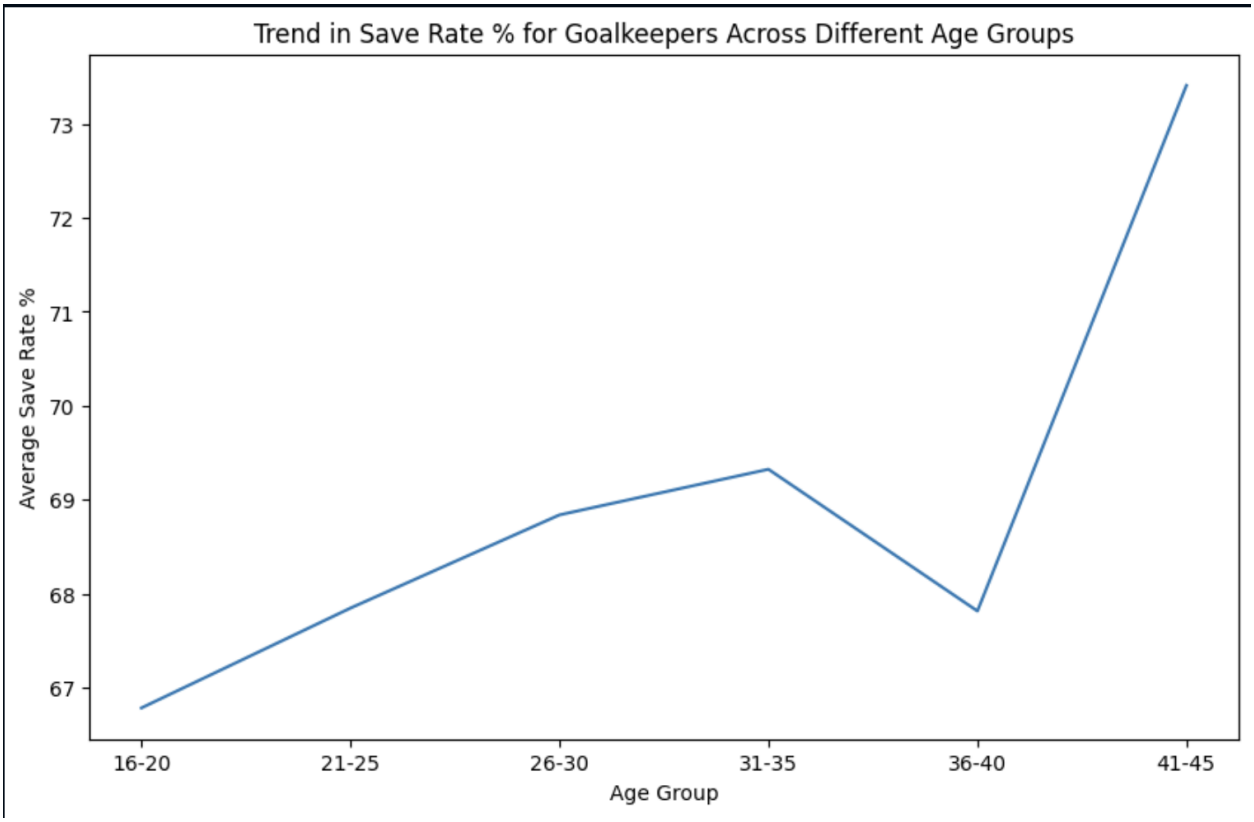
(Duels Won% and Aerial Duels Won% by Position)



(Relationship between Goals per 90 and Assists per 90 for Forwards)



(Relationship between Passes per 90 and Accurate Passes % for Midfielders)



(Trend in Save Rate % for Goalkeepers Across Different Age Groups)

3 Machine Learning

1.1.9. Method

The Gradient Boosting Classifier was the main focus for this project on creating a model for predicting goals. This machine learning algorithm belongs to the ensemble learning category and usually employs decision trees as its base learners. Its goal is progressively improving accuracy by rectifying previous errors in the tree sequence. The choice of Gradient Boosting Classifier was based on its ability to handle complex datasets effectively and its resistance to overfitting. This was deemed essential in predicting the goal-scoring potential of football players as football data is complicated and includes non-linear relationships between various performance indicators. The Gradient Boosting Classifier excels in discovering and utilizing these complex patterns, essential for making precise predictions in this multifaceted domain.

Moreover, the classifier's robustness against overfitting is critical in predictive modelling, especially in sports analytics. It is essential that the model not only memorizes the training data but also generalizes well to new, unseen data. With its sequential error correction and gradient optimization techniques, the Gradient Boosting Classifier effectively avoids overfitting. This ensures that predictions about a player's goal-scoring potential are reliable and applicable to different players, enhancing the model's usefulness in real-world situations. Therefore, these characteristics make the Gradient Boosting Classifier the most appropriate choice for this project.

1.1.10. Model evaluation

The model's evaluation was carried out using three key metrics, namely precision, recall, and the F1-score. Precision was prioritized to ensure that the model accurately identified goal scorers and minimized false positives. Recall was included to assess the model's ability to identify all potential goal scorers, which is crucial for player scouting. The F1-score, which combines precision and recall, provided a comprehensive measure of the model's overall effectiveness. To maintain the model's consistency and reliability, cross-validation techniques were applied across various data segments. This approach was instrumental in confirming the model's strength in different scenarios without any biases.

1.1.11. Model improvement

During the model improvement stage, the critical hyperparameters of the Gradient Boosting Classifier were fine-tuned. The adjusted hyperparameters included n_estimators, learning_rate, max_depth, min_samples_split, and min_samples_leaf. Each hyperparameter played a significant role in the model's learning process. For instance, n_estimators determined the number of trees in the model, directly affecting its complexity and potential for overfitting. Similarly, the learning rate influenced the contribution of each tree, where a lower rate typically required more trees but could lead to more accurate generalization.

RandomizedSearchCV, a technique that samples from a range of values for each hyperparameter was used to optimize these hyperparameters. This approach helped balance thoroughly exploring the hyperparameter space and maintaining computational efficiency. One main challenge was ensuring the model stayed balanced while retaining its generalization capabilities. Adjustments in hyperparameters like max_depth and min_sam-

ples_split were crucial in achieving this balance. For example, increasing max_depth allowed the model to capture more complex patterns and increased the overfitting risk. On the other hand, a higher min_samples_split value might simplify the model excessively, leading to underfitting.

The hyperparameters were carefully tuned to enhance the model's precision, recall, and F1 scores. These improvements indicated the model's enhanced predictive accuracy and strengthened ability to generalize across different datasets.

Overall, the machine learning aspect of the project entailed a careful selection of an appropriate model, thorough evaluation through established metrics, and strategic hyperparameter tuning to achieve optimal performance.

4 Ethical Considerations

Several ethical considerations emerge in the context of NAC's use of player data for analytics and developing a predictive model for goal scoring. These considerations align with the three vital elements of an ethical organizational capacity: respect for individual rights, transparency, and accountability.

1.1.12. Respect for Individual Rights

NAC must ensure that the collection and use of player data respect the players' individual rights. This includes adhering to privacy laws such as the GDPR in the European Union, which grants players control over their personal data. Players should have the right to access, rectify, and, in some cases, delete their personal data. Additionally, the club must obtain explicit consent from players before using their data, especially for purposes beyond performance enhancement, like commercial uses.

1.1.13. Transparency

Transparency in the use of player data is crucial. NAC must be transparent about what data is being collected, how it is used, and for what purposes. Players should be fully informed about the data analytics methods employed and how they might affect them positively and negatively. This transparency extends to the use of any machine learning models, where the criteria and algorithms used should be as clear as possible to avoid any form of bias or unfair treatment.

1.1.14. Accountability

According to ethical standards, the club should be accountable for the ethical use of data and should have policies and procedures in place to address any concerns or complaints from players regarding data misuse or privacy breaches. The club's management, especially those in charge of data analytics and player development, should be responsible for upholding these ethical standards to ensure that players' rights are protected.

In relation to NAC's handling of ethical elements, it is apparent that the club is aware of its responsibilities to protect players' rights. The privacy statement available on their website reflects their commitment to data protection and privacy. However, balancing the club's data-driven approach to team improvement with individual players' rights remains a complex challenge. The club needs to navigate this carefully, ensuring that their data practices do not compromise the players' autonomy or personal data security.

To ensure ethical decision-making, a structured framework was followed that included assessing potential impacts, considering alternatives, and consulting relevant guidelines. This approach aligns with the Ethical Guidelines for Statistical Practice and GDPR compliance.

Firstly, a thorough evaluation was conducted to assess how the use of player data could impact individual players, especially regarding privacy and data rights. Secondly, GDPR guidelines were followed to ensure that data handling practices respect players' privacy rights and consent. The Ethical Guidelines for Statistical Practice were also consulted to maintain integrity and fairness in data analysis. Lastly, alternative methods and models were explored to balance the club's objectives with ethical obligations, ensuring minimal privacy intrusion.

1.1.15. Ethical issues within NAC include:

- Data Privacy and Consent: There is a risk of violating players' privacy rights, mainly if their data is used without explicit consent or for purposes beyond what they agreed to.

- Transparency and Bias: The use of complex machine learning models may result in a lack of transparency in player evaluation, and potential biases in these models could lead to unfair treatment of players.

These issues require careful management to maintain ethical standards while using data analytics to enhance the team's performance.

1.1.16. To improve ethical guidelines, NAC should consider the following recommendations:

- Enhanced Consent Process: Implement a more robust consent process for players regarding data usage. This should include clear explanations of how data is used and the option for players to opt-out.
- Transparency in Data Analysis: Increase transparency around using machine learning models. This involves explaining to players how these models work and how they impact decision-making.
- Regular Ethical Audits: Conduct regular audits to ensure ongoing compliance with GDPR and ethical standards.
- Bias Monitoring in Models: Continuously monitor and test models for biases, ensuring fair treatment of all players.
- Data Usage Policy: Develop a comprehensive data usage policy that covers all aspects of data handling, including storage, access, and sharing.
- Training and Awareness: Provide training for staff on GDPR compliance and ethical data handling practices.

5 Recommendations

Here are recommendations for improving player recruitment for NAC Breda based on the findings from the exploratory data analysis and the machine learning model:

1.1.17. Strategic Player Profiling:
Develop detailed player profiles using the model. Focus on players whose attributes align with the model's indicators for high goal-scoring potential. This targeted approach can streamline recruitment efforts.

1.1.18. Expanding Data Sources:
Integrate data from various sources, including international leagues and lower divisions, to identify undervalued talent. This can provide a competitive edge in finding promising players before they attract wider attention.

1.1.19. Developing a Player Potential Index:
Create an index or scoring system to rank players based on various metrics identified as significant in the model. This index can simplify comparisons between potential recruits.

1.1.20. Collaborative Decision-Making Process:
Encourage collaboration between data analysts and scouts. Combining data-driven insights with traditional scouting expertise can lead to more nuanced assessments of players.

1.1.21. Incorporating Injury and Health Data:
Integrate injury history and health data into the recruitment process. This can help assess risks associated with potential signings and ensure long-term player availability.

1.1.22. Investing in Technology:
Invest in advanced data analytics tools and technologies to enhance the depth and speed of analysis. This could include AI-driven tools for real-time data analysis and visualization.

1.1.23. Customized Recruitment Strategies:
Tailor recruitment strategies based on the model's insights. For instance, if the model highlights the importance of a specific attribute like agility or endurance, focus scouting efforts on leagues or regions where players excel in these areas.

1.1.24. Performance Benchmarking:
Regularly benchmark the team's performance against key metrics highlighted by the model. This can help track the impact of recruitment decisions over time and guide future strategies.

By following these recommendations, NAC Breda can significantly improve its player recruitment process by leveraging a balanced approach, combining data-driven insights and traditional scouting expertise.

References:

Breda, N. (n.d.-b). *NAC Breda*. NAC Breda.

Netherlands Enterprise Agency, RVO. (2023, December 29). *10 steps to comply with the GDPR in the Netherlands*. business.gov.nl.

Chambers, R. (2021, 15 juli). State of the football analytics industry in 2021. SciSports. https://www.scisports.com/state-ofthe-football-analytics-industry-in2021/#:~:text=Technological%20advancements%20in%20football%20have,different%20personas%20within%20the%20game

Football Psychology. (z.d.). Google Books. https://books.google.nl/books?hl=en&lr=&id=F1OWDwAAQBAJ&oi=fnd&pg=PA297&dq=NAC+Breda+ethics+data++football&ots=M_rbAUXEu&sig=2N243rm1pnqomp0gVSPcX1r_Fa8&redir_esc=y#v=onepage&q&f=false

Herberger, T. A., & Litke, C. (2021). The Impact of big data and sports analytics on Professional Football: A Systematic Literature review. In Springer proceedings in business and economics (pp. 147–171).

Hummel, P., Braun, M., & Dabrock, P. (2020b). Own data? Ethical reflections on data ownership. Philosophy & Technology, 34(3), 545–572.

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