Advancing Talent Identification in Football through Machine Learning

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University OF APPLIED SCIENCES

Talent Prediction Perfection: Applying Machine Learning to Enhance Business Decisions for NAC Breda

Maria Salop

Applied Data Science & Artificial Intelligence, Breda University of Applied Sciences

Myrthe Buckens

June 21, 2024

Index

Talent Prediction Perfection: Applying Machine Learning to Enhance Business Decisions for NAC Breda

In the realm of professional football, the capacity to recognise and procure skilled players is crucial, not only for achieving success on the pitch, but also for ensuring long-term financial viability. NAC Breda, like other progressive football teams, is actively pursuing innovative approaches to improve its scouting and player assessment procedures. The main business problem is to efficiently and properly optimise the identification of talent and evaluate the market values of players while keeping costs low. This study seeks to address this problem by utilising Machine Learning.

The objective of this project is to create a predictive analytics framework that makes use of past data, player performance measures, and market trends to predict the future market value of a player. Our goal is to convert a wide range of numerical data into practical insights by utilising machine learning methods like linear and logistic regression. These valuable data will enable the club to make informed judgements when scouting players and negotiating contracts.

Furthermore, the research seeks to determine the key characteristics that have a substantial impact on a player's value in the market. This profound comprehension aims to reveal latent abilities in players who are frequently disregarded by conventional scouting techniques and identify prospects for strategic investments.

The report will tackle the business challenge by systematically going through the steps of data preparation, feature selection, model construction, and evaluation to ensure that our predictive models are strong and dependable. The objective is to create a scalable and adaptable machine learning framework that can effectively respond to the changing environment of football economics. This framework will provide NAC Breda a substantial advantage in the talent acquisition market.

1 Exploratory Data Analysis

The dataset includes 15,000 football players. Each of them was characterised by 134 distinct characteristics. The dataset contains a substantial amount of personal data, including basic information such as age, current team, weight, country of birth, and performance metrics. This comprises measurable information like number of matches played, minutes played, goals scored, and assists. Furthermore, additional detailed measurements such as expected goals, which offer a more profound understanding of a player's performance, are also incorporated.

1.1 Handling Missing Values

The initial data preparation stage involved evaluating missing values to prevent bias and impact machine learning models' effectiveness. I removed rows with over 18 missing values were to maintain study integrity. I also removed unrelated columns. I used Imputation procedures for missing columns like "Foot", "Age", and "Birth country", I replaced missing values with commonly occurring categories and median values to maintain central tendency.

1.2 Approaches for Dealing with Outliers

Outliers are crucial in data preparation and can distort results. I used statistical approaches like the Interquartile Range (IQR) technique, or Z-scores can help identify and remove abnormal data points. Domain-specific information can create unique rules for outlier detection, such as establishing criteria for unusually high or low player statistics in a football player dataset.

1.3 Data Transformation Techniques

Data transformation is a crucial step in preparing a dataset for machine learning algorithms. I used normalization or standardisation, adjusted numerical features to have an average of zero and a standard deviation of one. This is essential for algorithms like k-Nearest Neighbours and Support Vector Machines. I handled non-numeric data using one-hot encoding and label encoding, ensuring accurate interpretation without introducing ordinality. These transformations are crucial for models.

1.4 Data Descriptive Analysis

The dataset of football players' qualities includes statistical data on their age range, goals scored, and assists. The average age is 25.86, indicating a mix of youth and experience. The standard deviation is 4.7 years, indicating a diverse group of athletes. The dataset shows a wide range of scores, with both high and low scores present. The average assists per player is 3.26, reflecting various levels of team performance.

The average market value of players is 25.27 million euros, with a standard deviation of 13.26 million euros, ranging from 1.23 million euros to 49.42 million euros. The broad spectrum indicates a notable difference in how players are valued, influenced by things including skill, potential, and market demand.

1.5 Notable Patterns and Anomalies

The dataset reveals important patterns and anomalies, showing how the distribution of player positions affects market values and performance metrics, underscoring the importance of team dynamics in analyzing individual data. The wide range of market values highlights the economic aspects of football, influenced by factors beyond performance. This diversity underscores the complexity of

talent assessment, necessitating advanced models that integrate both quantitative and qualitative factors to accurately estimate a player's potential value.

1.6 Key Findings from Exploratory Analysis

The exploratory research conducted on the dataset to enhance talent identification in football using machine learning has provided significant insights. Significant trends, patterns, and theories have been revealed by a thorough analysis using various analytical and visual methodologies. The findings improve our understanding of what influences player performance and market value, and they also play a significant role in shaping the direction of future analytical and model development processes.

1.7 Identified Major Trends and Patterns

I noticed an intricate correlation between age and market worth, showing a rise in value as players mature up to a specific point, beyond which it may stabilise or decrease, demonstrating the balance between gaining experience and physical decline.

Performance measures like goals and assists showed a significant positive association with market value, highlighting the high value attributed to offensive abilities and their influence on a player's market value.

Differences in market values among player positions indicate a preference for forwards and attacking players, highlighting the importance placed on scoring goals.

The distribution of players among teams in the dataset indicated a possible sampling bias that may impact the applicability of predictive models.

1.8 Formulation of Hypotheses

Players in the peak of their careers may provide the highest value in terms of performance and potential for increased market value.

Emphasising players with strong goal and assist statistics can lead to substantial increases in both market value and on-field performance, underscoring the significance of these performance measures.

Player position valuation is dynamic and influenced by market developments, requiring adaptable algorithms for correct assessment.

1.9 Implications for Future Analysis

Feature Selection: I used correlations and patterns to strategically pick features for predictive modelling, emphasising the significance of age, goals, and assists as determinants of market value.

Data analysis requires meticulous data pretreatment and maybe stratified sampling to prevent bias from overrepresented groups influencing the models.

Model Selection: I observed linkages and complicated discovered patterns indicate the need to consider both linear and non-linear models to accurately capture the subtle influences on market valuation.

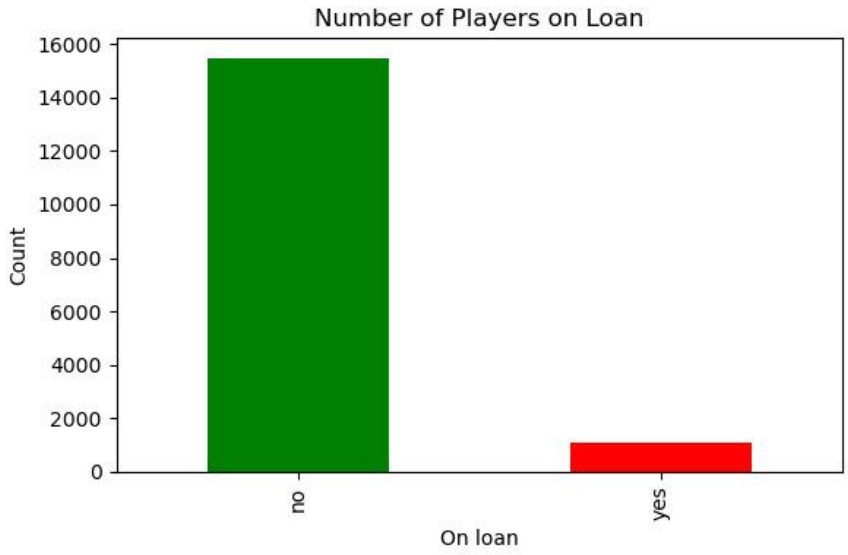

Figure 1: This Bar Chart displays the count of players categorised by their loan status. The majority of players, approximately 15,000, are not on loan ("no"), while a significantly smaller number, about 2,000, are currently on loan ("yes").

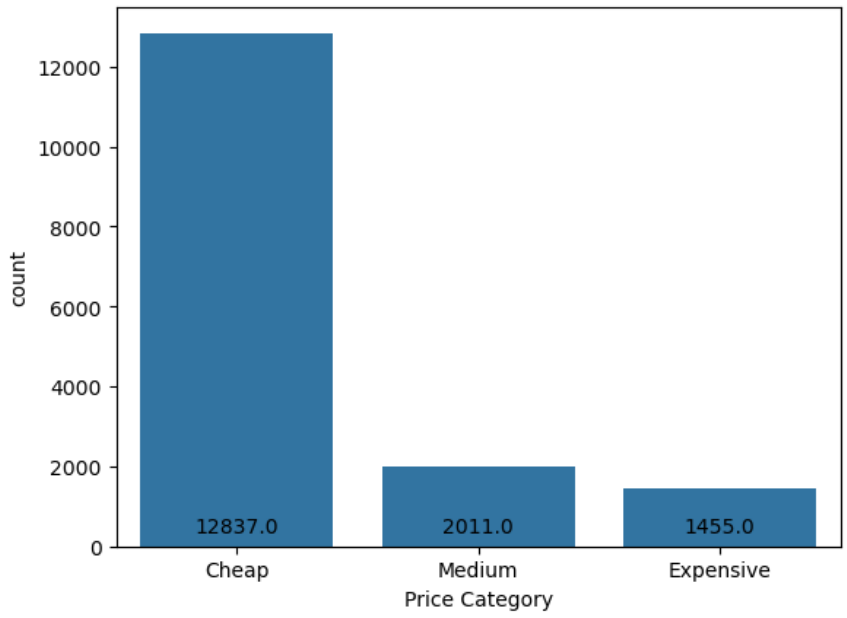

Figure 2: This bar chart illustrates the distribution of players across three price categories: cheap, medium, and expensive, based on their market value. The 'Cheap' category, which represents players

priced between €0 and €500,000, has the highest number of players, totaling 12,837. The 'Medium' category, with player prices ranging from €500,000 to €1,000,000, contains 2,011 players. The 'Expensive' category includes players priced over €1,000,000, with the fewest at 1,455. This visualisation highlights not only the relative abundance of players classified as cheap but also provides clear economic distinctions between each category.

2 Machine Learning

2.1 Method

I employed a Decision Tree Classifier for this assignment. This tool bears resemblance to a flowchart as it commences with a comprehensive inquiry and progressively segregates the data based on the responses, ultimately arriving at a definitive determination. This model is highly effective as it is capable of processing several forms of data, encompassing both numerical values and categorical information. This versatility is particularly advantageous for analysing data pertaining to football players.

The primary benefit of the Decision Tree algorithm is in its ability to provide comprehensive explanations for its judgements, explicitly demonstrating the rationale behind each data split. The clarity is of utmost importance for NAC Breda's scouting team as it enables them to have confidence in and comprehend the model's evaluation of prospective players.

2.2 Model Evaluation

I assessed the model's performance using two main methodologies: Accuracy and Mean Squared Error. Accuracy measures the proportion of accurate forecasts, providing a precise measure of the model's efficacy in categorizing data. The Mean Squared Error measures the average squared difference between actual results and model predictions. A smaller MSE indicates a higher level of precision and accuracy, indicating a closer match between the model's predictions and actual outcomes. These measures provide a comprehensive understanding of the model's reliability and precision, enabling informed decisions about its use in real-life situations.

2.3 Ensuring Reliability

I implemented the method of 5-fold cross-validation to evaluate the performance of the model. This approach entails partitioning the entire dataset into five subsets, sequentially utilising each subset to evaluate the model's performance while training it on the remaining subsets. This ensures the efficacy of our methodology not just in theoretical scenarios but also in practical, real-world applications.

2.4 Interpreting the Results

The Decision Tree model achieved an accuracy of approximately 78% and a mean squared error of 0.489. These statistics indicate that the model is highly dependable, however there is still potential for enhancement.

2.5 Model Improvement

The model underwent several adjustments by me to enhance its performance. I adjusted the tree's depth to determine the number of layers it should extend, affecting decision-making detail. I established minimum samples for splits were to ensure sufficient data points before further division, enhancing decision-making reliability. I defined minimal samples for leaves to prevent overspecificity and loss of generalizability. I carefully chose the splitting to ensure the most informative features are used for decision-making. These modifications are crucial for improving the model's precision and dependability, enabling it to capture fundamental data patterns.

These modifications facilitate the model's efficient learning process without excessive complexity or overfitting to our particular dataset. The equilibrium is vital for constructing a model that not only exhibits high performance with our existing data but also demonstrates adaptability to novel, unobserved data.

3 Ethical Considerations

3.1 Ethical Considerations for AI in NAC Breda

NAC Breda must integrate ethical values when using dashboard analytics to evaluate player performance. This paper explores the ethical implications of such technology, specifically addressing data privacy and informed consent, transparency and fair play, and accountability and continuous monitoring. The document identifies responsible parties and evaluates NAC's adherence to standards like GDPR and the Ethical Guidelines for Statistical Practice.

3.2 Ethical Company

NAC Breda is a corporation that adheres to ethical principles and goes above the requirements of GDPR to ensure fairness and confidentiality in managing player performance data. The company also prioritises diversity in treating employees, maintains openness in its operations, and values employee input. This comprehensive strategy enhances confidence and ensures the company's long-term sustainability, fostering trust and social engagement.

3.3 Ethical Process and Tools

NAC Breda places a high importance on ethical integration in their sports analytics AI technologies, with a focus on guaranteeing openness and fairness in the usage and analysis of data. Systematic audits are regularly performed to detect and avoid any biases that may lead to biassed ratings of players. The company's dedication to technology utilisation is demonstrated by its proactive approach to ethical issues in both the creation and deployment of AI technologies.

3.4 Ethical Conduct of Employees and Clients

The ethical framework of NAC Breda is clearly demonstrated by the conduct of its professionals towards stakeholders, including as players, coaching staff, sponsors, and supporters. They uphold stringent norms of professional behaviour, comprehend the ethical ramifications, and demonstrate responsible behaviour. Providing employee training and ongoing education on ethics guarantees that all individuals adhere to the company's ethical standards, while also cultivating a culture characterised by respect and appreciation.

3.5 Data privacy and informed consent

NAC Breda prioritises data privacy through robust security measures, including advanced encryption, rigorous access controls, and secure storage systems. Comprehensive consent forms outline data collection purposes, usage, and potential consequences, ensuring players' informed consent and

promoting transparency. This practice not only protects players' personal information but also fosters trust between the club and its athletes.

3.6 Potential Ethical Problem: Bias in Data Interpretation and Decision Making

NAC Breda faces a challenge with potential biases in its AI-driven player analytics. These biases might lead to unfair assessments, affecting players' careers and potentially breaching the club's ethical standards of fairness and equal treatment. This issue could undermine trust in the analytics system.

3.7 Recommendations for Ethical Enhancement

1. Update Data Protection Technologies: Regularly updating technologies to enhance data protection, as well as conducting training sessions to educate players on data security practices.
2. Enhance Transparency: Clearly explaining evaluation metrics and their implications to players and other stakeholders to build trust.
3. Strengthen Accountability: Defining explicit roles for ethical oversight and establishing protocols for addressing ethical violations promptly and effectively.

3.8 Conclusion

Adhering to key ethical principles—data privacy and informed consent, transparency and fair play, and accountability and continuous monitoring—is crucial for NAC Breda. This commitment will foster trust and enhance the club's reputation for ethical AI utilization. By continuously improving its ethical standards and practices, NAC Breda can set an example for other organisations in the sports industry.

4 Recommendations

The following concise recommendations are provided to NAC Breda to enhance talent discovery and value, while maintaining ethical standards in the use of machine learning and analytics.

4.1 Improving Machine Learning Models

To enhance player evaluations, NAC Breda should diversify performance metrics and external factors in data collection, apply sophisticated machine learning algorithms for better predictions, and regularly update models to maintain relevance. Additionally, advanced encryption and governance should be implemented to boost data privacy. Transparent communication of evaluation criteria will improve clarity for all stakeholders, and forming an Ethical Oversight Committee will help monitor and assess ethical compliance effectively.

4.2Continuous improvement and innovation

NAC Breda should provide training in data science and ethical AI to foster a culture of learning and innovation among employees. Engaging in collaborative research projects with academic and tech partners can explore innovative and ethical sports analytics methods. Additionally, monitoring technological trends will help maintain cutting-edge practices in sports analytics.

4.3 Promotion of fairness and equity

Perform bias audits periodically on machine learning models to ensure fairness in player evaluations.

Implement inclusive data practices by creating data collection methods that appropriately represent player diversity and reduce the dangers of prejudice.

NAC Breda may improve its talent identification and valuation accuracy, uphold ethical integrity in analytics use, and promote a culture of continual innovation and justice by following these streamlined guidelines.

References

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. In M. Peitz & J. Waldfogel (Eds.), *Handbook of Media Economics* (pp. 441-499). Springer. https://link.springer.com/chapter/10.1007/978-3-030-27994-3

EIDH. (2023). NAC Breda: New Member. Retrieved May 20, 2024, from https://www.eidh.org/blog/member/nac-bredal

GDPR.eu. (n.d.). Consent under GDPR. Retrieved May 20, 2024, from https://gdpr-info.eu/issues/consent/

GDPR.eu. (n.d.). What is personal data?. Retrieved May 20, 2024, from https://gdpr-info.eu/issues/personal-data/

NAC. (n.d.). De Club. Retrieved May 20, 2024, from https://www.nac.nl/de-club

The text was reviewed and paraphrased using QuillBot (https://quillbot.com).

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