

Optimizing Football: Predictive Player Position Modelling

Unlocking Potential: Predictive Analytics in Football



DISCOVER YOUR WORLD

Index

Contents

Index 1		
1	Introduction	3
1.1	Business problem	3
1.1.1	Optimizing Football Player Positions	3
1.1.2	Balancing Risk and Data in Player Positioning	3
1.1.3	Predictive Models for Player Performance	3
1.1.4	Revolutionizing Player Position Allocation	3
1.2	Solution	4
1.2.1	Transforming Player Positioning with Advanced Predictive Modelling	4
1.2.2	Predictive Analysis for Optimal Player Positions	4
1.2.3	Data-Driven Evolution in Football Strategy	4
2	Exploratory Data Analysis	5
2.1	Introduction:	5
2.2	High-Level Overview of the Dataset:	5
2.3	Steps Taken to Prepare the Data:	5
2.4	Visual Techniques:	5
2.5	Examining Relationships Between Variables:	6
2.6	Summary of Key Findings:	6
2.7	Conclusion:	6
3	Machine Learning	7
3.1	Method	7
3.1.1	Employing Various Machine Learning Models	7
3.1.2	Model Selection and Optimization	7
3.2	Model evaluation	7
3.2.1	Evaluating Model Performance	7
3.2.2	Metrics Selection	7
3.2.3	Insights from Model Performance	7
3.3	Model improvement	7
3.3.1	Fine-Tuning with Grid Search	7
3.3.2	Automating Optimal Parameter Selection for Random Forest Classifier	8
3.3.3	Feature Selection and Model Development	8
4	Ethical Considerations	9
4.1	The three elements vital for an ethical organizational capacity to NAC:	9
4.2	Identify the responsible parties within NAC for each of these elements:	9
4.3	Research on how NAC takes the ethical elements into consideration:	9
4.4	Provide proof that I followed at least one framework for making ethical decisions:	9
4.4.1	GDPR Considerations:	9
4.4.2	Ethical Guidelines for Statistical Practice:	10
4.5	Ethical problems I have identified within NAC:	10
4.6	Recommendations for NAC to improve their ethical guidelines:	10
5	Recommendations	11
5.1	Looking for players:	11

5.2	Analysing scouted players:	11
5.3	Consideration for Player Acquisition:	11
6	Referencing	12
6.1	Reference the use of a Generative AI program:	12
6.2	Reference webpage(s) used in my report:	12
6.3	Scholarly sources:	12

1 Introduction

1.1 Business problem

1.1.1 Optimizing Football Player Positions

Football clubs worldwide grapple with a pervasive challenge—accurately predicting the optimal positions for their players. This complex task poses a formidable hurdle, hindering clubs from fully optimizing team performance and making informed decisions in player recruitment and development. The crux of the matter lies in the necessity of placing players in positions that allow them to unleash their maximum potential. Consider the scenario where a club invests substantially in a player, only to witness lacklustre performances in the subsequent season. The root cause often traces back to the player being played in a wrong position at the new club, underscoring the critical importance of strategic position allocation.

1.1.2 Balancing Risk and Data in Player Positioning

Existing datasets provide valuable insights into positions where a player has previously excelled. However, clubs exhibit a natural tendency towards caution and risk aversion, hesitating to experiment with new positions based solely on historical data. In a season filled with weekly matches, clubs cannot afford to jeopardize a player's performance during crucial games. The fear of losing points or matches due to experimenting with player positions becomes a substantial deterrent, emphasizing the need for a proactive and risk-mitigating solution.

1.1.3 Predictive Models for Player Performance

The pressing need for football clubs is a predictive model capable of anticipating a player's performance in various positions without subjecting them to real-time testing during high stakes matches. Avoiding this risky trial-and-error approach is not only crucial for maintaining consistency but also for sustaining competitiveness throughout the season. The quest for such a model is underscored by the desire to make strategic decisions that transcend conventional wisdom, leading to enhanced team dynamics and improved player performance.

1.1.4 Revolutionizing Player Position Allocation

The absence of a dependable predictive model continues to impede clubs' ability to allocate resources effectively and tap into the full potential of each player on the field. It represents a critical juncture where innovation can offer a transformative solution. This is where my idea takes centre stage—a solution poised to revolutionize how football clubs approach player position allocation, ultimately reshaping the landscape of player development and team strategy.

1.2 Solution

1.2.1 Transforming Player Positioning with Advanced Predictive Modelling

A groundbreaking solution to this pervasive challenge involves the implementation of an advanced predictive modelling system for player position allocation. Leveraging the power of cutting-edge machine learning algorithms, this system delves into a diverse array of player-specific features. It goes beyond conventional metrics, considering physical attributes, playing style nuances, historical performance data, and positional tendencies. By synthesizing this wealth of information, football clubs can develop a robust and adaptable model capable of predicting the most suitable position for each player with decent accuracy.

1.2.2 Predictive Analysis for Optimal Player Positions

The overarching goal of my analysis is to predict, based on comprehensive and nuanced statistics, whether a player would excel as a forward, midfielder, defender, or goalkeeper. This holistic approach not only streamlines decision-making for clubs but also empowers them to make data-driven choices that enhance team dynamics and player performance across various positions. The model acts as a strategic compass, guiding clubs towards optimal player utilization and mitigating the risks associated with positional experimentation.

1.2.3 Data-Driven Evolution in Football Strategy

In essence, the proposed solution represents a paradigm shift in how football clubs approach player development and strategic decision-making. It aligns with the broader industry trend of embracing data-driven insights to gain a competitive edge. The implementation of such a predictive model is not merely a technological upgrade but a strategic imperative for clubs aiming to stay ahead in the ever-evolving landscape of professional football.

2 Exploratory Data Analysis

2.1 Introduction:

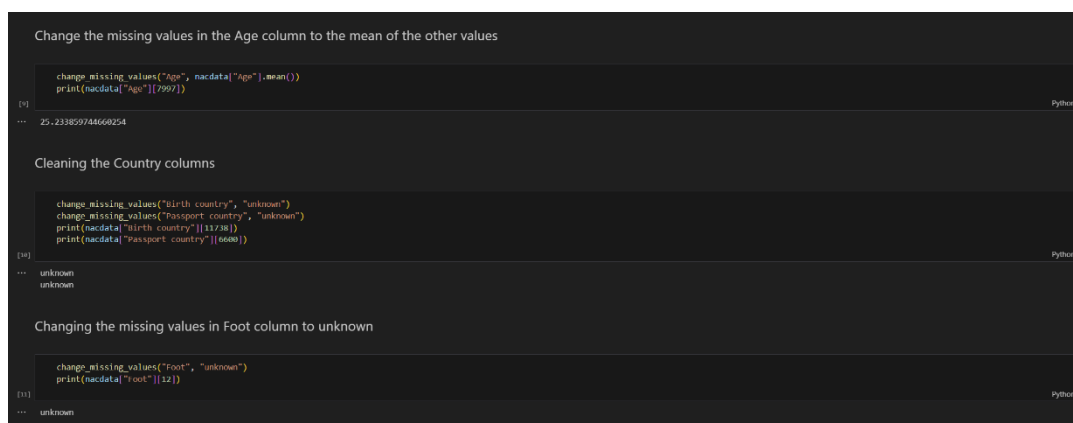
The overarching aim of my analysis is to harness predictive modelling to anticipate player positions, encompassing forwarders, midfielders, defenders, and goalkeepers, based on a diverse set of features. This analytical endeavour is strategically designed to serve football clubs by offering a predictive tool that enhances their ability to optimize player performance and make well-informed decisions regarding the allocation of players to specific positions on the field.

2.2 High-Level Overview of the Dataset:

The consolidation of the dataset involved a meticulous process of merging information from diverse sources, resulting in a dataset of substantial magnitude. It encompasses a notable number of records and features, drawing information from the provided datasets. These features include both numerical attributes such as Age, Weight, and Height, and categorical attributes like Player positions, Birth country, and On loan status. The pre-processing steps undertaken were multifaceted, involving the removal of NaN values, categorization of features, and filling numerical values with appropriate averages or 0 values.

2.3 Steps Taken to Prepare the Data:

The data preparation process unfolded over an extended period, involving systematic removal of NaN values and standardization of categorical features by replacing missing values with 'unknown.' Numerical features, including Age, Weight, and Height, underwent a refinement process where missing values were filled with averages. Other numerical features like Head goals, Free kicks per 90, and penalties taken were addressed by filling with 0 values. Data transformation techniques, such as normalization, standardization, and encoding, were implemented, with specific adjustments made to columns for an enriched analytical approach.



```
Change the missing values in the Age column to the mean of the other values

change_missing_values("Age", nacdatal["Age"].mean())
print(nacdatal["Age"][[729]])

[74]
... 25.233859744660254

Cleaning the Country columns

change_missing_values("Birth country", "unknown")
change_missing_values("Passport country", "unknown")
print(nacdatal["Birth country"][[117:18]])
print(nacdatal["Passport country"][[6600]])

[74]
... unknown
... unknown

Changing the missing values in Foot column to unknown

change_missing_values("Foot", "unknown")
print(nacdatal["Foot"][[12]])

[74]
... unknown
```

Figure 1: Steps Taken to Prepare the Data (CTRL + Click on the picture to reach it on GitHub)

2.4 Visual Techniques:

A suite of visualizations, including histograms, scatter plots, and box plots, was strategically employed to delve deeper into data distribution, identify outliers, and discern patterns. These visual representations played a pivotal role in influencing decisions related to feature selection and illuminating potential relationships within the dataset. The nuanced insights derived from visual techniques contributed significantly to the overall data analysis strategy.

2.5 Examining Relationships Between Variables:

The exploration of relationships between variables was facilitated through the calculation of correlation coefficients for numerical features. This analytical approach shed light on the intricate connections within various facets of player statistics. The implications of significant correlations or their absence were carefully considered, providing valuable context for understanding the dynamic interplay between different variables and guiding subsequent analytical decisions.

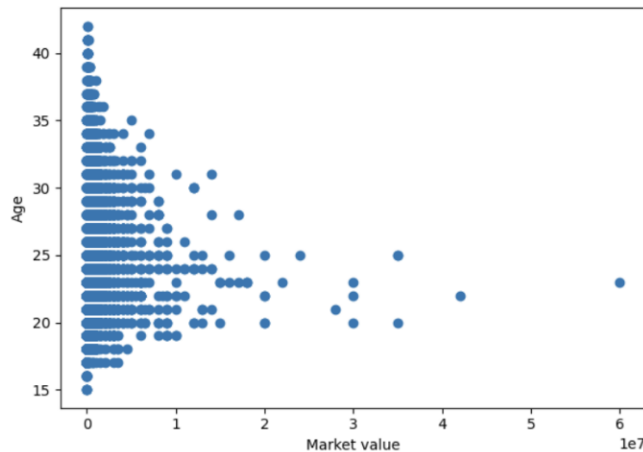


Figure 2: Correlation between players age and their market value (CTRL + Click on the picture to reach it on GitHub)

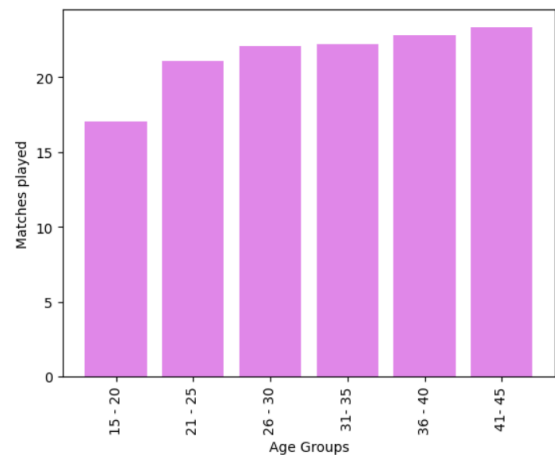


Figure 3: Average number of matches played by different age groups (CTRL + Click on the picture to reach it on GitHub)

2.6 Summary of Key Findings:

The culmination of the analysis revealed major trends and patterns within the dataset, offering profound insights into the distribution of player positions and related statistics. These findings provided a fertile ground for the formulation of potential hypotheses, laying the groundwork for further in-depth investigation. The identified trends and patterns hold considerable significance in shaping subsequent phases of data analysis, influencing feature selection, and informing the strategic choice of predictive models.

2.7 Conclusion:

In conclusion, the amalgamation of diverse datasets has yielded a wealth of valuable insights. The scrupulously prepared dataset serves as a robust foundation for the forthcoming predictive modelling phase, poised to empower football clubs in their pursuit of optimizing player performance and making judicious decisions regarding player positions on the field. The journey from dataset consolidation to meticulous preparation has set the stage for an informed and data-driven approach to enhancing team dynamics and achieving strategic success.

3 Machine Learning

3.1 Method

3.1.1 Employing Various Machine Learning Models

In the methodology employed for this analysis, a range of machine learning models was systematically applied, including Regression Tree, Linear Regression, Logistic Regression, Gradient Boosting Tree, and Lasso Regression.

3.1.2 Model Selection and Optimization

The rationale behind selecting these models stemmed from their compatibility with the dataset, providing a foundation for seamless data fitting. Early on, I had envisioned a set of potentially beneficial algorithms for future use, and these were among the ones considered. Ultimately, the Random Forest Classifier emerged as the most effective model, delivering an impressive accuracy of 88.5% when optimized with the best parameters, a result that left me thoroughly satisfied.

3.2 Model evaluation

3.2.1 Evaluating Model Performance

To comprehensively evaluate the performance of the chosen model, a suite of metrics was employed. This included the implementation of a K Nearest Neighbour Classifier, a visualization through the Plot Confusion Matrix, and a detailed Classification Report. The report encompassed key metrics such as precision, recall, f1-score, support, accuracy, macro average, and weighted average.

3.2.2 Metrics Selection

These metrics were chosen based on their clarity and their ability to provide nuanced insights into the model's performance over weeks of analysis.

3.2.3 Insights from Model Performance

The overall model performance stood out with an impressive 86% accuracy. A closer examination through the classification report highlighted the model's proficiency in accurately predicting goalkeepers, while defenders exhibited a high accuracy as well. Challenges surfaced in distinguishing between forwards and midfielders, owing to their similarities in various measures. This observation implies that these two positions hold significant potential for improvement and further refinement.

3.3 Model improvement

3.3.1 Fine-Tuning with Grid Search

The pursuit of an optimal model led to the adoption of the Grid Search technique, aimed at fine-tuning the Random Forest Classifier. The appeal of Grid Search lay in its ability to seamlessly integrate with the Random Forest Classifier, allowing for the identification of the best parameters, ultimately yielding an accuracy of 88%.

3.3.2 Automating Optimal Parameter Selection for Random Forest Classifier

Overcoming the challenge of steering the Random Forest Classifier towards the best estimator was a significant accomplishment, achieved through the automation of the classifier to follow the optimal parameters.

3.3.3 Feature Selection and Model Development

The result, maintaining an accuracy level around 88%, demonstrated consistency with the expectations set at the beginning. Despite the proximity to the initial Logistic Regression result, the meticulous selection of the most relevant features from the outset played a pivotal role in achieving and maintaining this high level of accuracy throughout the model development process.

4 Ethical Considerations

4.1 The three elements vital for an ethical organizational capacity to NAC:

Data Privacy: NAC should prioritize protecting sensitive information accessed through the dashboard, ensuring data security, and promoting a culture of privacy awareness among staff and players.

Data Quality: NAC should follow ethical principles to ensure accurate, valid, and reliable data collection, representing the target population without biases.

Fairness: NAC should strive for fairness in dashboard use, ensuring algorithms and data analyses are free from discrimination.

4.2 Identify the responsible parties within NAC for each of these elements:

Data Privacy: The responsibility lies with the IT and security teams, as well as management to promote a culture of data privacy.

Data Quality: Data Scientists and Analysts play a crucial role, along with management overseeing adherence to ethical guidelines.

Fairness: Data Scientists and Analysts are responsible for ensuring fairness in algorithms and analyses.

4.3 Research on how NAC takes the ethical elements into consideration:

NAC actively considers data privacy, data quality, and fairness in its use of dashboards, emphasizing the importance of these ethical elements in maintaining trust and integrity. No trace for discrimination in their dashboards, and within their sponsorship partners.

4.4 Provide proof that I followed at least one framework for making ethical decisions:

In my opinion the emphasis on data privacy, data quality, fairness, transparency, and user rights align with general ethical frameworks for data usage and decision-making.

4.4.1 GDPR Considerations:

Data Privacy: My findings emphasize the importance of data privacy, stating that NAC should prioritize the protection of sensitive information, implement comprehensive data security measures, and promote a culture of data privacy awareness.

User Rights: NAC should respect the rights of individuals whose data is collected, including obtaining informed consent, providing individuals with control over their data, and honouring requests for data access, correction, or deletion.

4.4.2 Ethical Guidelines for Statistical Practice:

Data Quality: My research discusses ethical considerations in dashboards to ensure accurate, valid, and reliable data. This aligns with the ethical guidelines for statistical practice, which emphasize the importance of unbiased, accurate data collection and interpretation.

Transparency: NAC must ensure transparency in how data is collected, analysed, and used within the dashboard system. Transparency is a key aspect of ethical statistical practice, allowing stakeholders to understand the sources of data, methodologies used, and criteria for decision-making based on the insights derived from the data.

Fairness: NAC should strive for fairness in their use of the dashboard, ensuring that algorithms and data analyses are free from discrimination. Fairness is a fundamental ethical principle in statistical practice, ensuring unbiased and equitable treatment of individuals or groups in data analysis.

4.5 Ethical problems I have identified within NAC:

While prioritizing data protection and implementing comprehensive security measures is undoubtedly important, the extent of these measures might be excessive or financially burdensome for the club. This emphasizes that NAC should primarily focus on its core activities related to training and competitions. The allocation of extensive resources towards data security might divert funds and attention away from these primary objectives.

Ethical considerations, while essential, might not entirely eradicate biases or ensure the absolute accuracy and representativeness of data collected through dashboards for NAC. One aspect of this could emphasize the inherent limitations in eliminating biases entirely. Despite adhering to ethical principles, biases can persist in data collection due to various factors, including the nature of the data sources, the methodology employed, or inherent societal biases that are challenging to identify and eliminate completely. Human interpretation and decision-making involved in data collection and analysis can inadvertently introduce biases, even with stringent ethical guidelines in place. Another aspect to consider is that while ethical principles guide responsible data collection, they might not address external factors that can influence data integrity, such as unforeseen events, changing societal norms, or evolving technologies. These external factors could potentially impact the accuracy and representativeness of data despite ethical adherence.

4.6 Recommendations for NAC to improve their ethical guidelines:

Balancing Resources: Find a balance between data security and core activities to avoid diverting excessive funds and attention.

Continuous Improvement: Acknowledge inherent limitations in data collection and interpretation and strive for continuous improvement.

Training and Communication: Emphasize ongoing training for staff and players on ethical guidelines and maintain open communication to address concerns.

External Factors: Develop strategies to address external factors that may influence data integrity despite ethical adherence.

5 Recommendations

5.1 Looking for players:

In the quest for new talent, it's worth exploring players who have showcased excellence in a particular season, boasting impressive stats and witnessing a surge in market value, only to experience a subsequent decline. To facilitate this search, tapping into comprehensive datasets that meticulously document players' market values over the years becomes crucial. Platforms like [transfermarkt.com](https://www.transfermarkt.com), renowned for their expansive football player data, can be a valuable ally in this endeavour.

5.2 Analysing scouted players:

Following the identification of potential players, a more in-depth analysis comes into play. Leveraging a specialized model, a thorough examination can be conducted to determine whether a player's current positions align with the anticipated performance metrics for their respective roles. This involves a comprehensive assessment of various key metrics to ensure that the player is strategically positioned to contribute optimally to the team's dynamics.

5.3 Consideration for Player Acquisition:

As the analysis phase concludes, a pivotal decision-making juncture emerges. If there's a disparity between the predicted position and the actual performance values, it presents a unique opportunity. In such instances, serious consideration should be given to acquiring the player. The rationale lies in the potential to enhance the player's overall performance by strategically placing them in positions identified by the predictive model. This strategic approach not only optimizes the player's strengths but also positions the team for improved overall performance and success.

6 Referencing

6.1 Reference the use of a Generative AI program:

- OpenAI. *ChatGPT*. The report was done, then I wanted a creative title and subtitle for it. Prompt: "Give me a title of a report written about a predictive machine learning model on football data", (24-01-2024)
- OpenAI. *ChatGPT*. Source text was written by me and then expanded. Prompt: "Please broaden the following text:", (22-01-2024)
- OpenAI. *ChatGPT*. Source text was written by me and then I wanted to make it to look more professional. Prompt: "Make this text to a little bit more professional:", (22-01-2024)
-

6.2 Reference webpage(s) used in my report:

- Transfermarkt. (n.d.). *Football transfers, rumours, market values and news*.
<https://www.transfermarkt.com/>

6.3 Scholarly sources:

- Velasquez, M. G. (2017). *Business Ethics: Concepts and Cases*. Pearson.
- Twin, A. (2023, March 17). *Business Ethics: Definition, Principles, Why They're Important*. Investopedia.
<https://www.investopedia.com/terms/b/business-ethics.asp>
- *5 Principles of data Ethics for business*. (2021, March 16). Business Insights Blog.
<https://online.hbs.edu/blog/post/data-ethics>



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