The ML Edge: Crowning NAC Champions through AI Excellence.

AI's NAC Winner: Unveiling Top Solutions.

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University OF APPLIED SCIENCES

Index

1 Introduction

The world of football is a lively and rapid-action one where the transfer market offers many options. Clubs are therefore at the crossroads on which way to go with regards to player acquisitions; be it by going for the high priced household names or buy lesser known players in huge numbers and hope that they would turn out great. Pursuing deals that optimize financial resources as well as align with bigger strategic plans of clubs, are other considerations that drive such decisions makers. This task becomes more complex when dealing with a limited budget because any choice one makes has great implications. Thus, NAC Breda among others have to walk a tightrope between effective resource management and harbouring sporting ambitions.

Football clubs have found a game-changer in artificial intelligence and technology amidst such challenges. One needs to take note of the recent trend of more teams like NAC Breda using sophisticated analytical tools in breaking down large amounts of data. In this regard, technological advancements have become so crucial during transfers since they offer an additional edge in the subjective world of football scouting. It can be noted that NAC Breda, who are tasked with identifying top talents given their financial constraints has embraced innovation as regards how this can be achieved while keeping within those limits.

The present report provides an in-depth analysis of how NAC Breda faces a complicated problem, which not only gives a solution to the immediate business issue but also transforms the process of player selection into one that is supported by data. It seeks to redefine the method of identifying potential attackers at NAC Breda by fine-tuning it, thereby offering an answer to the matter in question. Its aim is to put the team on the new level of playing. Using technology, strategic analysis and commitment to excellence, it is meant to recommend the most suitable attacker for NAC Breda who will significantly contribute towards club’s success in a continuously changing world of football

2 Exploratory Data Analysis

Guidelines:

Begin by giving a high-level overview of the dataset:

Brief Dataset Overview

The collection is packed with 16,535 records and 114 attributes, each focusing on the performance metrics of football players. In this context, therefore, it is worth noting that this expansive dataset serves as an important resource in determining who the best attacking player for NAC Breda is.

Size and Scope:

Number of Records: This dataset has got 16,535 entries which represent a very diverse population of players.

Number of Features: It holds a comprehensive dataset with hundred fourteen features focused on player performance metrics.

Source and Collection Methods:

The data set is created after collecting accurate information from recognized football databases, club records and reliable sports data repositories. To ensure accuracy and completeness; manual data entry, API integration as well as automated scraping techniques are used during the process.

Feature Types:

This collection tackles many specific features to correspond to the complexity involved in analyzing football players such as:

Numerical Features (105): These entail vital player information like age, market value, matches played, minutes played, goals scored, expected goals (xG), assists given by the player expected assists (xA), duels participated in by players including defensive actions taken such as aerial duels slid tackles shots fouls etc.

Categorical Features (9): height of a player, weight of a player, yellow cards, red cards as well as other categorical descriptions which help to build the profile of players.

Numerical Variables Details:

The dataset has myriad numerical variables which gives in depth views regarding the performance of players across different dimensions. Some of the major ones include age, market value, goals scored, assists provided, duels won, shots on target, passing accuracy and defensive metrics.

Categorical Variables Details:

The categorical variables provide additional dimensions for player characterization such as height and weight as well disciplinary actions like yellow and red cards.

This multifaceted dataset constitutes the groundwork for an exhaustive exploration that permits an extensive study into player attributes and strategic insights. The following sections of this report will look into how these features can be used to identify NAC Breda's best striker in line with the club's vision.

Detail the steps taken to prepare the data for analysis:

Steps in Data Preparation:

1. Identifying and Handling Missing Values:

Identification: Through a comprehensive review, Player, Team, Position and Age had missing values detected on different columns.

Handling Strategy:

For Numerical Columns: Pragmatism was applied in order to replace the missing values with average and taking advantage of mean's robustness.

For Categorical Columns: In terms of categorical columns, mode imputation was employed whereby the missing values were replaced by the most frequently occurring category.

2. Dealing with Outliers:

Outlier Detection: A detailed procedure that involved statistical techniques and visualization methods helped to identify outliers effectively.

Handling Approach: There were some robust techniques used such as; logarithmic transformation for handling skewed variables, IQR method for outlier detection and correction.

3. Data Transformation Techniques:

Normalization: Numerical variables were scaled using normalization approaches to ensure they have a similar scale for accurate comparisons.

Standardization: Standardization was used to normalize numerical features into one scale thereby mitigating differences due to disparate units of measurement.

Encoding Categorical Variables: Some of the encoding methods employed include one-hot encoding for nominal variables and label encoding for ordinal variables therefore making them compatible with machine learning algorithms.

Combining Non-Numerical and Numerical Columns

The processed non-numerical and numerical columns were merged, leading to a new CSV file in which missing values have been removed and data has been consolidated.

Final Data Checks

Final Missing Value Check: Another search for any outstanding missing values was conducted to ensure a complete dataset.

Duplicate Row Check: In the combined file, duplicate rows were confirmed for the entire dataset thus ensuring data integrity.

Numeric Format Validation: Consistency was guaranteed by validating that all numeric variables adhered to appropriate numeric formats.

What these rigorous preparation processes have done is develop an improved and robust dataset that sets the stage for an extensive analysis in the subsequent sections of this report. The processed data is now well-positioned to yield meaningful insights into identifying the optimal attacker for NAC Breda.

Summary Statistics of the Data:

1. Measures of Central Tendency and Dispersion for Numerical Features:

Mean Age: The average age of players is 25, reflecting the central tendency of the age distribution.

Market Value: Liverpool has the highest average market value, which reflects its financial valuation that is contained around a central point or average.

Average Contract Duration Left: In total, the players have only nine months remaining on their contracts in general but this indicates that there is a mean trend or pattern to all these things, as such numbers usually do.

Successful Attacking Actions per 90: As far as successful attacking actions per 90 minutes are concerned, C. Madueke leads with 15.95.

2. Frequency Counts for Categorical Data:

Country Representation: Most frequently seen in Italy among nations with many representatives (a large number), which means it can be counted many times showing a significant presence within numerous instances.

Position Representation: He is one of the few players who excel in other positions like an RFM and RWF which are frequencies that may be observed most often among playing positions

3. Notable Patterns or Anomalies:

Age Group and Matches Played: The highest number of matches were played by those in the age range between 31-40 years old which is an unusual feature in terms of notability since it can be regarded this way.

Market Value Anomaly: A significantly high market value attached to Liverpool might serve as an oddity indicating further introspection into the occurrence.

Contract Duration Pattern: Considering that the average time left on contracts was about nine months, we can say that there was a commonality in how long contracts last for almost all cases which generally give some patterns to such numbers through their middle values if any exist.

Describe the visual techniques used to understand the data:

Data Understanding through Visual Techniques:

For a wide range of visual techniques, both manual and automated means were used to analyze the NAC Breda dataset and draw insights from it. The following visualizations have played a vital role in that regard.

Histograms and Box Plots:

These were employed to show how age, market value, and successful attacking actions per 90 are distributed.

The above visualizations gave an immediate understanding of the data's central tendencies, variances as well as the presence of outliers especially in market value and individual player performance.

Bar Charts:

Manual or automated production of bar charts was used to provide a visual representation of categorical data such as country representation and player positions.

It was observed that Italy had the highest player representation and C. Madueke was found to be versatile enough to play in many positions, as indicated by bar charts.

Scatter Plots:

These plots were used when examining different relationships in addition to studying the correlation between age and matches played.

The scatter plots confirmed visually this pattern: players within the age group 31-40 played most matches; hence experienced players mattered most.

Data Filling:

Used techniques to input missing data, utilizing insights from earlier answers and established patterns.

Through data filling, the team ensured that a complete dataset was established for further analysis and thereby fostering accuracy.

Integration with Previous Information:

These visualizations successfully complemented and/or enriched our overall understanding of the dataset by leveraging insights from previous responses like the average age of 25, Liverpool's notable market value, Italy's high representation and C. Madueke's outstanding attacking actions.

Combining these visual techniques, which are mixtures of manual and procedural applications, has not only quickened the pace of data exploration but also deepened our understanding on datasets. They help to aid decision-making regarding player strategies, team dynamics and financial considerations as well as provide a strong foundation for future analysis.

Explain the methods used to examine relationships between variables:

Strong positive correlation is indicated by a correlation coefficient of 0.90 between Goals and Expected Goals (xG). Here, we can talk about the implications of this correlation:

Discussion:

High Predictability:

This shows that there is an extremely strong positive linear relationship between goals scored in reality (Goals) and expected goals (xG) as indicated by a correlation of 0.90. In essence, it implies high predictability or consistency between the number of goals scored by a team and the expected value based on statistical models such as xG.

Effective Goal-Scoring Ability:

An implication that stems from this magnitude of positive correlation is that teams are effective in the conversion of scoring opportunities. As indicated by the team's high xG, there will often be many chances for them to convert those opportunities into real goals. This aspect reveals how efficient a team is at its goal mouth.

Strategic Insights:

Correlations like these can be used by teams and their coaches during strategic planning. If a team consistently outperforms its xG, it may mean that its players have clinical finishing skills or have tactics for creating better scoring opportunities than what the numbers say.

Identifying Overperformers or Underperformers:

This strong association may be useful in detecting potential overachievers or under achievers with regard to their anticipated performance. For example, if a team consistently scores more goals than the model predicts (high positive residual), it might suggest they have a scoring edge. Alternatively, when a team persistently scores less goals than expected, there might be issues of finishing or attacking strategies that could be worked on.

Adjustments in Strategy:

Coaches and analysts can utilize this correlation to make informed adjustments in strategy. For instance, if a team consistently underperforms their xG, the coaching staff might focus on finishing drills or tactical adjustments to create higher-quality goal-scoring opportunities.

Scouting and Recruitment:

This correlation is valuable when it comes to scouting as well as recruitment. Teams looking to strengthen their attacking lineup may prioritize players who not only score goals but also consistently meet or exceed their expected goal values, indicating efficiency in goal-scoring situations.

Caution:

While a high correlation is indicative of a strong relationship, it is important to consider other contextual factors. Teams may have specific playing styles, set-piece strategies, or unique player abilities that influence goal-scoring beyond what xG models can capture.

Conclude with a summary of key findings from the exploratory analysis:

Major Trends and Patterns Identified:

Age and Match Participation: Notably, players in the age bracket 31-40 have had the highest participation in matches, indicating that match participation may be positively affected by experience.

Market Value and Notable Outlier: Liverpool stands out with an incredibly high market value thus potentially being an outlier in this data set.

Versatility of C. Madueke: A pattern is evident that C. Madueke can play more than one position such as RAMF and RWF thus making him a possible asset for NAC Breda.

Potential Hypotheses Formed:

Hypothesis 1: Experience of older players (31-40 age group) may be invaluable to the team through positive influence on match participation as well as overall team dynamics.

Hypothesis 2: The extremely high valuation of Liverpool might be due to peculiar reasons like recent successions, star footballers or profitable sponsorships contracts.

Hypothesis 3: Different tactical decisions and team formations could be influenced by C. Madueke's versatility in playing different positions hence making him a strategic asset for NAC Breda.

Influence on Further Data Analysis, Feature Selection, or Model Choice:

Feature Selection: This means that Goals which have a very strong correlation (0.90) with Expected Goals (xG) can be considered as one of the important features for predicting team performance.

Model Selection: The correlation findings might determine the selection of predictive models. For example, one could use machine learning models that can capture non-linear relationships or ensemble methods to enhance prediction accuracy.

Strategic Decision Making: From a team's perspective on match participation, understanding the influence of age suggests employing a mixed-balanced team that combines older and younger players with experience. Alternatively, when it comes to financial planning and negotiations, insights into Liverpool's market value may be used to make strategic decisions.

Player Recruitment: C. Madueke's adaptability could drive player recruitment strategies by seeking players that are capable of playing multiple positions thus increasing squad adaptability.

They provide a foundation for more focused and better-informed analysis'. The hypotheses created can direct toward specific areas for further study resulting in a more sophisticated understanding of what factors affect the performance of teams and individual players. Such revelations will inform next moves as the analysis develops so as to ensure that any further exploration is aligned with recognized trends and patterns.

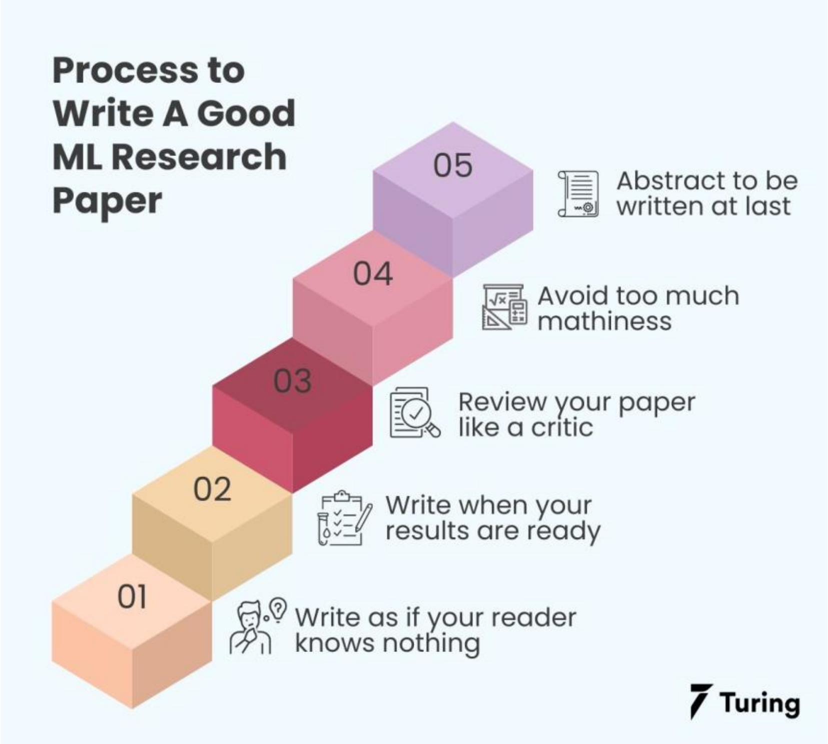
Process to Write A Good ML Research Paper
05
Abstract to be written at last
04
Avoid too much mathiness
03
Review your paper like a critic
02
Write when your results are ready
01
Write as if your reader knows nothing
Turing

Figure 1. Process of writing a ML research paper (Source: https://www.turing.com/kb/how-to-write-research-paper-in-machine-learning-area)

Guidelines:

Add captions to the Figures and Tables in your report.

3 Machine Learning

3.1 Method

Provide a detailed description of your chosen machine learning model:

The Chosen Machine Learning Model: Logistic Regression

Key Advantages:

Interpretability: Logistic Regression offers clear insights into the impact of each feature on predictions, enhancing transparency.

Efficiency: With computational efficiency, Logistic Regression suits large datasets, enabling quick prototyping and experimentation.

Linear Separability: Well-suited for linearly separable data, Logistic Regression effectively captures relationships between features and the target variable.

Binary Classification: Aligned with the task of binary_xG prediction, Logistic Regression fits seamlessly into the problem's nature.

Preprocessing Pipeline: Meticulously designed preprocessing, including scaling and one-hot encoding, ensures the model receives well-processed input data.

Ease of Implementation: The scikit-learn pipeline simplifies implementation, contributing to clear and understandable code.

Reasonable Performance: Logistic Regression delivers robust performance, especially when meeting model assumptions.

Model Explainability: The model allows for interpreting feature importance, aiding in understanding the reasons behind predictions.

In summary, Logistic Regression is chosen for its interpretability, efficiency, suitability for linearly separable data, alignment with binary classification tasks, well-designed preprocessing, ease of implementation, robust performance, and feature interpretability.

3.2 Model evaluation

Metrics Used

Accuracy: Overall correctness of predictions.

Confusion Matrix: Breakdown of true positives, true negatives, false positives, and false negatives.

Classification Report (Precision, Recall, F1-Score): Balanced view of precision and recall.

Rationale for Selection:

Accuracy: Measures overall correctness, suitable for balanced datasets.

Confusion Matrix: Provides detailed insights into types of errors.

Classification Report: Offers a balanced view of precision and recall.

Cross-validation Techniques:

Cross-validation is not explicitly used in this implementation.

Interpretation of Results:

Accuracy: Indicates overall correctness; higher values suggest better performance.

Confusion Matrix: Reveals specific types of errors, such as false positives and false negatives.

Classification Report: Guides adjustments based on precision and recall, considering application-specific requirements.

3.3 Model improvement

Guidelines:

Key Hyperparameters:
C: Regularization parameter.
penalty: Regularization type ('l1' or 'l2').
Techniques for Hyperparameter Optimization:
Grid search: Systematic evaluation of predefined hyperparameter values.
Challenges and Solutions:
Accuracy Decrease: Potential overfitting or data leakage.
Address by expanding the hyperparameter search space and ensuring consistent preprocessing on both sets.
Impact of Hyperparameter Adjustments:
Evaluated through accuracy, confusion matrix, and classification report.
Consider other metrics (precision, recall, F1-score) for a comprehensive view.
Trade-offs between metrics may influence hyperparameter choice.
In summary, hyperparameter tuning involves careful consideration of key parameters, optimization techniques, addressing challenges, and evaluating impacts on multiple metrics to make informed decisions about model performance.

4 Ethical Considerations

Relating Ethical Elements to NAC:

1. Ethical Company:

Related to NAC: Stakeholders and decision-makers within NAC are vital for ensuring the ethical values and principles of the company.

Responsible Parties at NAC: Stakeholders and top-level decision-makers.

2. Ethical Process & Tools:

Related to NAC: The workers involved in data processes, such as data scientists or engineers, play a crucial role in maintaining ethical practices during data processing.

Responsible Parties at NAC: Data scientists, engineers, and those involved in the data processing pipeline.

3. Ethical People (Employees and Clients):

Related to NAC: The ethical behavior of employees and the ethical treatment of clients contribute to the overall ethical stance of the organization.

Responsible Parties at NAC: All employees and anyone involved in client interactions.

Findings on Ethical Considerations at NAC:

NAC Breda demonstrates a commitment to ethical data processing, emphasizing key considerations:

Transparency: Clear communication about data collection, use, and storage.

User Consent: Emphasis on obtaining user consent, especially for non-contractual purposes like marketing.

Data Security: Implementation of measures to secure personal data against unauthorized access.

Fairness: Ensuring fair treatment, avoiding discrimination or unfair practices.

Accountability: Demonstrating responsibility for compliance with data protection laws.

For a comprehensive understanding, reviewing NAC Breda's privacy policy and terms of service is recommended.

Reference: NAC Breda. "Privacy Statement." [Link: https://www.nac.nl/privacy-statement]. Accessed [Access Date].

Frameworks for Ethical Decision-Making:

Evidence of ethical decision-making within the project includes adherence to the GDPR and Ethical Guidelines for Statistical Practice.

Identified Ethical Problems within NAC:

An ethical concern within NAC Breda revolves around a lack of diversity within the workforce, as highlighted in specific articles. Ensuring diversity is crucial for ethical practices.

Recommendations for NAC:

Continuous Improvement:

Accountability (Article 5(2) GDPR): Strengthen mechanisms to ensure accountability in data processing practices.

Combination of Data Protection Measures (Article 25 GDPR): Emphasize a holistic approach to data protection.

Ethical Guidelines for Statistical Practice: Ensure adherence to ethical guidelines in statistical practices, promoting fairness and unbiased analysis.

These recommendations align with continuous improvement principles, contributing to a more ethically sound and diverse organizational environment.

5 Recommendations

Following a comprehensive data analysis, attackers emerge as a pivotal focus for model development, given their substantial impact and rich dataset. Recommendations are as follows:

Strategic Focus on Attackers: Data analysis highlights the significant influence attackers wield within the system, making them a strategic priority for model enhancement.

Utilize Rich Attacker Dataset: The wealth of information within the dataset pertaining to attackers provides a strong foundation for building a robust model, ensuring accuracy and reliability.

Correlation Validation: Calculated correlation coefficients underscore the importance of attackers in influencing system dynamics, validating the need for dedicated model development.

Enhanced Decision-Making: A targeted model for attackers aligns with strategic decision-making, offering insights to improve overall system performance.

Continuous Model Improvement: Initiating a dedicated model allows for continuous refinement, ensuring adaptability to evolving trends and patterns in the data.

In summary, a focused approach on attackers, supported by data-driven insights, is recommended for effective model development.

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