Quantifying football Player Goal and Performance

A Predictive Analysis Using Market Value, Assists, Shots per 90, and Minutes Played and Goals per 90

Carcea Madalina no.221989

DISCOVER YOUR WORLD

2023-24Y1B ADS_AI /Department of
Data Science and Artificial Intelligence

Friday 26 January 2024

Breda University OF APPLIED SCIENCES

Appendix

1 Introduction

Football is a constantly evolving sport that requires teams to stay ahead of the curve in terms of player recruitment and overall team development. To achieve this, clubs must have a deep understanding of the factors that impact a player's ability to score goals. Our client, NAC Breda, is one such club that recognized the importance of gaining insights into the predictive elements associated with a player's goal-scoring capabilities.

To address this need, we utilized advanced analytics and machine learning techniques to build a predictive model that examines the complex relationship between a player's market value, assists, shots per 90 minutes, total minutes played, goals per 90 minutes, and goal-scoring proficiency. By delving into the intricate details of player statistics, we aimed to provide our clients with actionable insights that facilitate informed decision-making in terms of player investments, tactical planning, and overall team strategy.

Our predictive model provides a detailed analysis of individual player performance metrics and their collective impact on goal outcomes. We analysed each player's market value, assists, shots per 90 minutes, total minutes played, goals per 90 minutes, and goal-scoring proficiency to determine which factors were most significant in predicting a player's ability to score goals. By doing so, we aimed to provide our clients with a better understanding of the complex relationship between these variables and how they influence a player's goal-scoring capabilities.

Our report aims to empower our client with a data-driven approach that optimizes their ability to identify and nurture goal-scoring talent within their football organization. By providing our clients with actionable insights based on our predictive model, we aim to help them make informed decisions that will improve their overall team strategy, player recruitment, and tactical planning.

2 Exploratory Data Analysis

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

The dataset we are working with, called NAC, contains a total of 16535 records. Initially, it had 114 features. In order to keep the original dataset unaltered, I created a copy of it for data cleaning and exploratory data analysis (EDA). The new version of the dataset has 16527 rows and 115 columns after being cleaned. The copy created for EDA purposes has the same dimensions as the cleaned version. We received all the necessary files from NAC Breda to compile the complete dataset (NAC BREDA [1]).

At the beginning of the project, the data contained 4 types of data: objects, floats, integers and datetime, which were used to configure our business idea.

Detail the steps taken to prepare the data for analysis:

To handle missing values in my dataset, I employed the isnull() function to detect them, and the dropna() function to remove the values from the column. Subsequently, I used the Simple Imputer from the scikit-learn library to fill the gaps in the data. Additionally, I utilized Winsorization to manage any outliers detected during my analysis. For numerical columns, I opted to use standardization as a data transformation technique, which was implemented using the Standard Scaler. These are the approaches I utilized to deal with missing values and outliers in my dataset.

Provide summary statistics of the data:

The following summary statistics were calculated for numerical features such as 'Age', 'Market value', 'Assists', 'Shots per 90', and 'Minutes played'. These statistics include mean, median (50%), variance and standard deviation. Notable patterns were identified during exploratory data analysis. It is good to see that our data doesn't have any missing values. However, due to the large number of positions, it is difficult to visualize this data in any type of graphs.

Different types of plots, such as box plots, histograms, scatter plots, heatmaps, and pair plots, are used as visual techniques to understand data. The histogram displays numerical features, such as the column "Age," to determine if age plays a significant role in the sport. The scatter plot is utilized to identify relationships and potential correlations or patterns among the columns "Age," "Goals," and "Assists." The box plot is designed to reveal any outliers in the selected columns, which may require further investigation to solve any issues discovered. The visualizations offer valuable insights that can inform subsequent analysis steps. For example, any outliers detected through visualizations may necessitate handling techniques like winsorization. Moreover, correlation patterns observed in scatter plots can influence the choice of features for modeling. Lastly, comprehending feature distributions assists in making decisions about data transformation techniques, including normalization or standardization.

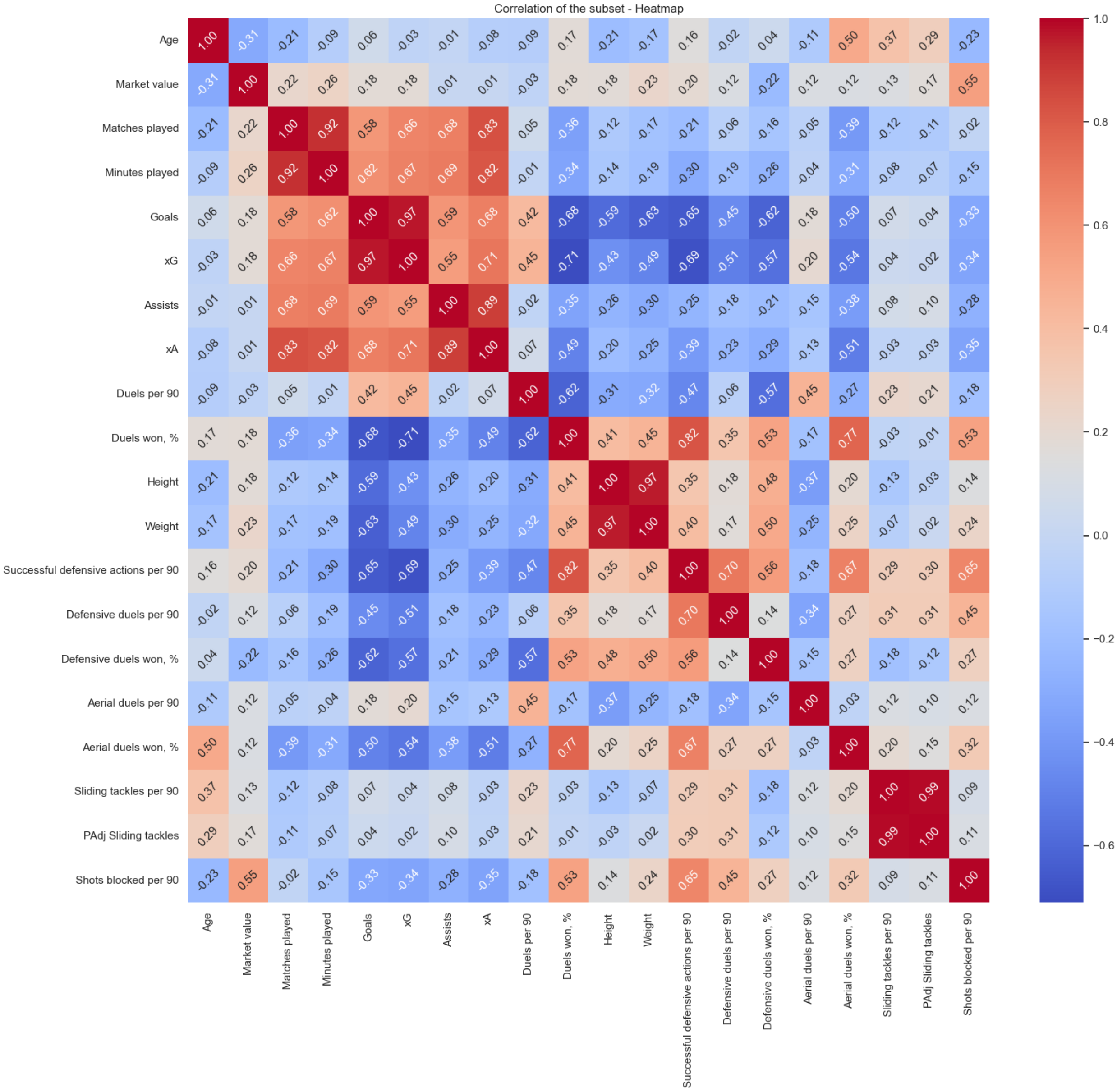
Explain the methods used to examine relationships between variables:

To investigate the relationship between variables, I used the Correlation Coefficients for numerical features. After running the code block, I observed a mix of both positive and negative correlations, but the variables ultimately exhibited high correlations. For instance, there was a strong positive correlation between the number of Matches played and Minutes played (correlation coefficient ≈ 0.88), indicating that players who participate in more matches tend to accumulate more playing time. Additionally, the correlation between 'Penalties taken' and 'Goals' was notably high (correlation coefficient ≈ 0.60), suggesting that players who take more penalties tend to score more goals.

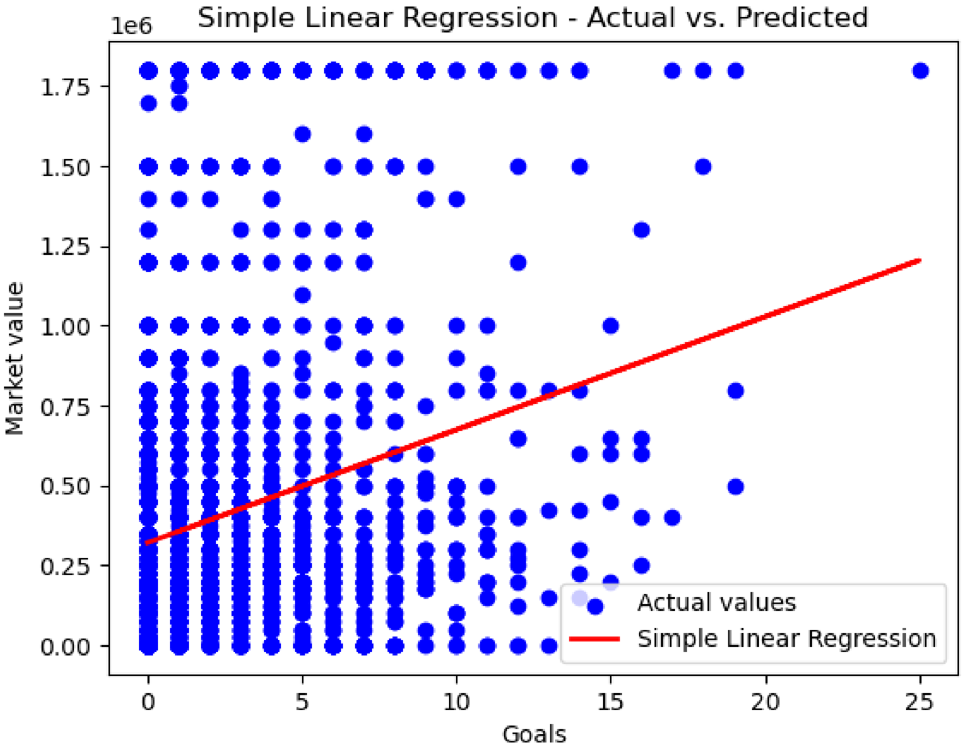
Conclude with a summary of key findings from the exploratory analysis:

One major trend observed is that there is a strong positive correlation between "Goals" and "Assists". This suggests that players who perform well in one area generally perform well in the other as well. Additionally, it was found that a player's "Market value" has positive correlations with various performance metrics, indicating that the value of a player in the market reflects their overall contributions. Moreover, a mild positive correlation exists between "Age" and "Market value," implying that older players may command higher market values.
As a potential hypothesis I will choose the positive correlation between Age and Market value, which suggests that, experienced players may demand higher market values doe to their skills, knowledge and expertise.

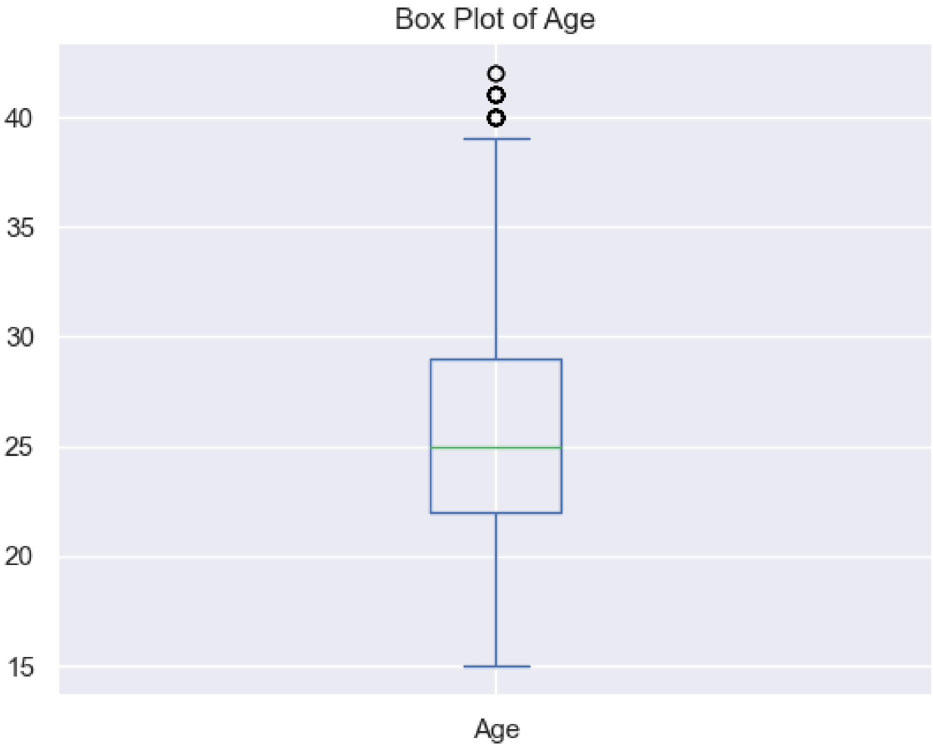
These findings offer valuable insights that can guide subsequent stages of analysis, feature selection, model development, and a more in-depth exploration of player age dynamics. At the feature selection stage, I based my model training, by considering the impact of the players age and the strong correlation between Goals, Assists and Minutes Played. These findings influenced my choice of model.



Correlation matrix between 20 features on X-axes and Y-axes



Linear regression on the model



Box plot of age to see which age is more frequently stored in the data set

3 Machine Learning

3.1 Method

Provide a detailed description of your chosen machine learning model:

The following machine learning algorithms were utilized:

- *Random Forest Regression and Classification*: I picked this specific algorithm as it can handle both regression and classification tasks. That was useful in finding the desired model and providing robust predictions by aggregating the outputs or warnings/errors. This algorithm captures complex relationships and provides insights into feature importance.

- *Linear Regression*: This simple yet effective model was used to establish a baseline for regression tasks and understand the linear relationships between features and the target variable. This is showcased in the previous pages, from the added visualizations.

- *Logistic Regression*: I used this model for classification tasks, which is suitable for binary or multiclass classification problems, which I encountered thus provides probabilities of class membership.

- *Gradient Boosting (Regressor)*: I explored this interesting method to capture intricate patterns and relationships. It builds decision trees step by step, with each tree correcting the errors of the previous one, ultimately helping with the accuracy score.

- *K-Means Clustering*: I applied K-Means for unsupervised learning to identify potential clusters or patterns within the data. It helps to explore groupings that might exist among football players based on selected features.

Other techniques used include pipelines to streamline the workflow by combining preprocessing steps and model training, hyperparameter tuning using *GridSearchCV* to fine-tune model parameters for optimal performance and *learning curves* and *cross-validation* to assess model performance, detect *overfitting*, and guide further refinement.

3.2 Model evaluation

Explain the methods used to evaluate the model's performance:

To assess the performance of the model, I utilized regression metrics including mean squared error, r-squared, and mean absolute error which helped me in measuring the accuracy of predictions and the variance in the model. For classification models, I used accuracy, precision, recall, F1 score, and confusion matrixes to evaluate the overall performance in terms of positive predicted value, balanced precision and recall, and other such measures. Additionally, I calculated mean squared, and r-squared values from the gradient boosting model for assessing the accuracy of predictions and the fit of the model.

The metrics were chosen to provide an overall performance measure for classification and insights into the predictive accuracy and goodness of fit for regression models.

I utilized the learning curve as a cross-validation technique. The regression model showed better performance, while gradient boosting indicated improved predictive accuracy. By combining these metrics with cross-validation, we were able to achieve the best performance for the model. This helped me get a better understanding of the decision-making process.

3.3 Model improvement

Describe the process of tuning and optimizing the model:

For hyperparameter optimization during tuning, I utilized techniques such as random search and grid search. My focus was primarily on the key parameters n_estimators, max_depth, min_samples, max_samples (for the RandomForestClassifier), and learning_rate and max_depth (for the GradientBoostingRegressor).

However, I faced some challenges during the tuning process. At one point, my model stopped working and started throwing errors, which forced me to switch from grid search to random search. I had to adjust my approach to ensure that I could complete the task at hand.

Despite these challenges, using hyperparameters resulted in a noticeable improvement in the model's performance. Specifically, there was an increase of 2-3% in accuracy.

4 Ethical Considerations

In this section, describe all your findings related to ethical considerations.

The three most vital elements for an ethical organization such as NAC are transparency, accountability, and integrity. Leaders, managers, and communication teams work together to maintain these three big elements.

In my opinion, the ability to find information about NAC Breda is crucial for transparency. This includes not only information about the football club, but also campaigns, staff, team members, and scandals. Being able to access the terms and conditions on the website along with the security policy is equally important.

Moreover, NAC Breda has shown accountability not only for their mistakes but also for their supporters' misconduct, which has led to police involvement and fines. This demonstrates that the company takes responsibility for its actions and is committed to maintaining ethical standards.

The integrity of NAC Breda is reflected in the good work they have done. This includes helping people become more active, giving opportunities to people in need or those who cannot afford sports, anti-racism(NAC Breda Geeft Racisme de Rode Kaart, n.d.) [3] campaigns (Breda, n.d.) [4], and educating young children who attend their under 20 teams.

Our project places a strong emphasis on ethics and responsible AI practices. To ensure ethical considerations are upheld at every stage of the project lifecycle, we conduct regular team discussions, carefully review documentation, and adhere to ethical principles during data collection, analysis, and model deployment.

To ensure the highest level of data protection, we strictly adhere to GDPR regulations. This means that all data handling processes are designed to meet General Data Protection Regulation standards.

We follow rigorous ethical guidelines for statistical analysis to ensure unbiased and transparent results that can be relied upon. In addition, we have integrated documentation of data anonymization, secure storage, and ethical statistical practices into our project workflow. These measures help to ensure that our project is both fair and safe.

I have noticed an ethical issue relating to the guest team changing room. The design has some flaws which give an advantage to NAC. They can easily change the temperature in the room to make the football players from the opposing team uncomfortable. Furthermore, there is a shortage of equipment and medical staff available to the guests. Even though this things count as tactics for the game, I don't consider them to be ethical, you either win because you were that great, or you lose ,but ultimately learn something that can help in the future , or similar circumstances.
It appears that there is a concern regarding the presence of Amstel and Unibet advertisements on their official website, despite the fact that children can access the site. Upon conducting research, it was discovered that to create an Unibet account, one must be at least 25 years old. However, there is always the possibility of someone lying about their age or using their parents' identity to create an account and debts.
Describe the ethical problems you have identified within NAC.

As a suggestion for NAC, I would like to enhance the ethical guidelines by increasing transparency and knowledge, particularly for players. For instance, it would be beneficial to know if all football players sign a contract agreeing to the collection of their data and statistics for future research purposes. End with recommendations for NAC to improve their ethical guidelines within the current project.

5 Recommendations

Write the recommendations to the client based on your findings for the problem statement in the Creative Brief.

Based on the analysis of the provided problem statement by NAC Breda, our team recommends that players goals and age be considered during the feature selection process as it can impact market value. We further suggest prioritizing features such as "Goals," "Assists," and "Minutes played" with significant correlations for predictive modelling. It is essential to take ethical considerations into account to ensure fairness and bias assessment during model deployment. Additionally, we recommend regularly updating the model with new data to maintain its relevance and accuracy.

6 Sources

1.Data Set Source: https://study.buas.nl/content/enforced/8758-FAI1.P2-01_2023/NAC%20Data.html?ou=8758&d2l_body_type=3

2. Breda, N. (2023, July 1). *Terms and Conditions* [Review of *Terms and Conditions*].

https://www.nac.nl/files/knvbstandaardvoorwaardenper1juli2023-def.pdf.

file:///D:/Downloads/knvbstandaardvoorwaardenper1juli2023-def%20en.pdf

3. *NAC Breda geeft racisme de rode kaart*. (n.d.). Rode Kaart. Retrieved November 30, 2023,

https://www.rodekaart.eu/news/nac-breda-geeft-racisme-de-rode-kaart/

4. Breda, N. A. C. (n.d.). Test. NAC Breda. https://www.nac.nl/identiteit

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