NAC Breda report

Louie Daans

DISCOVER YOUR WORLD

Breda University OF APPLIED SCIENCES

Index

1 Introduction

NAC Breda is a Dutch professional football club located in Breda evolving in Eredivisie, the first division of Netherlands. Their main goal, like every club out there, is to become stronger and thus win as many matches as possible. In modern football it is essential to use data science to compete with other clubs and improve performance consistently. Taking the role of a Data Scientist working for NAC Breda, I will do my best during this block to help the client improve their business processes using data, mathematics, and machine learning. I will more particularly focus on helping the club get insight on the provided dataset of football players, I will try to contribute to their player acquisition strategy by doing EDA, creating a market value predictor using machine learning and presenting ethical and legal challenges.

2 Exploratory Data Analysis

2.1 Dataset overview

The data used for this project comes from a folder containing 45 different excel files, where each file contains data about players from a specific league that NAC could potentially acquire. Since the files all have identical columns, I'm able to single source the data into a data frame named "data", which contains 16535 different players and 114 different features. Our dataset contains both numerical and categorical features, which will be analysed during EDA.

2.2 Preparing the dataset for EDA

2.2.1 Handling missing values

The first step is to handle the null-values in our dataset. The missing values could mean that the data is unknown, unavailable, not applicable, or could even be the result of a simple input error. Identifying the origin of those null values is effective since it will influence how we treat them. By creating a function that returns information about the null values in our dataset we can find 120 738 total null values distributed among 85 different columns of our dataset. In percentages, our dataset contains 6.35% null values. To ensure good analysis and machine learning compatibility, we must eliminate these values. I noticed that 232 players each have over 60 missing values, indicating an input error. Since players with more than half of their data missing useless to our EDA, we will drop them.

- The empty values present in the column 'Team' have a clear meaning. In this column we can see the club in which the player played in before joining the current team he plays in, thus for some young players it is normal to have empty values in 'Team' since they haven't played in a professional team previously, that's why we can fill in those values with 'none'.

- 5 null values are present in the column 'Age', since the amount is small, we can just replace them with the mean age.

- In the column 'Contract expires', more than half of the values are missing. It is recommended to delete such columns but here it seems to be meaning something. Since the expiration of contracts is critical to our analysis, in a way that knowing contract expiration or not could perhaps impact market value and other factors, I will replace empty values with 'unknown'.

- The column 'Birth country' contains 2 null values, for which I will just drop the rows containing them.

- In column 'Foot', I will replace empty values with 'both', assuming that players that are neither left-footed or right-footed are automatically both.

For the remaining columns containing null values, a clear pattern is observable where for six columns, the columns are only applicable to goalkeepers, and thus every non-goalkeeper player has missing values in them, and for the three other columns, they are not applicable to goalkeepers, and thus every goalkeeper has missing values in them. This means that we can just replace all the remaining null values with 0. We now have no more empty values left in our data.

2.2.2 Formatting values

Now that our dataset is free of empty values, we can focus on other abnormal values that we want to change. I first fixed some issues in the columns 'Weight', 'Height', 'Age' and 'Position'. But the most important issue lies in 'Market value'. 2548 players have a market value of 0. This could skew our EDA and later our prediction model. Accepting these values would imply that these players are worthless, which is not true. They simply lack a definitive market value estimation. To ensure accurate market value predictions, we should remove these players from the dataset. Dropping these 2548 players reduces the dataset but will result in having a more realistic project.

2.2.3 Creating additional columns

To improve analysis, I added three new columns to our dataset based on other columns information:

- The new column 'BMI' contains the body mass index of each player, which is used to express the relationship between height and weight as a single number and is often used to determine overweight and underweight.
- The new column 'Changed team' has two possible values; 'yes' if the player was in another team during previous season, 'no' if the player stayed in the same team.
- Finally, the new column 'General position' groups the possible 21 positions that players could play into 4 general positions: 'Goalkeeper', 'Defender', 'Midfielder' and 'Attacker'.

Our cleaned dataset of 13753 different players (rows) and 118 features (columns) is now ready for EDA.

2.3 Summary statistics of the data

Our dataset contains 106 numerical features which are mostly performance measures of players, and 12 categorical features that contain personal and professional information about each player.
This is the frequency of counts for data inside our categorical features:

Feature	Number of unique values	Examples
Player	13132	T. Biasci, J. Patricienne, P. Iemmello, ...
Team	1457	Ajax, Catanzaro, none, ...
Team within selected timeframe	763	Ajax, Catanzaro, OFK Pirin, ...
Position	21	CF, RB, GK, ...
Contract expires	77	2024-06-30, 2023-12-31, unknown, ...
Birth country	147	Italy, France, Germany, ...
Passport country	1035	Italy, France, Germany, ...
Foot	3	left, right or both
On loan	2	yes or no
League	43	Italy Serie C, Eredivisie, Iceland, ...
Changed team	2	Yes or No
General position	4	Attacker, Midfielder, Defender, or Goalkeeper

Table 1. Frequency of counts for data inside categorical features. (Source: NAC notebook)

Market value of players in our dataset ranges from 10,000 to 60,000,000, with an average market value of 610,580. A more balanced way to understand the distribution of market value is by finding the mean, which is 300,000.

Age ranges from 17 to 42 years old, with the average age of our players being 26.

From all the teams in our dataset, Ajax has the highest average market value of around 14.73 million per player (having 18 players in their team).

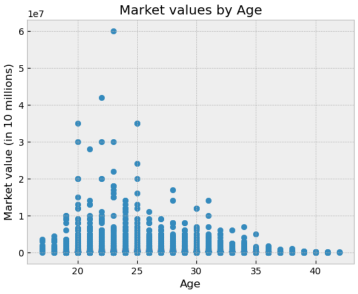
Italy has the highest representation in terms of player birthplace with 1588 players born there. Germany is second place with 1033 players and France third with 886 players.

I created a new column ‘xG_Diff’ in our dataset that contains the difference between expected and actual goals scored for each player. Then I searched for the top 10 players having the highest expected goals values and analysed their xG difference. In general, it is close to expectations, except for P.Iemmello that did way better than expected. This is useful to search for players that did way better than expected and could help make recruitment decisions.

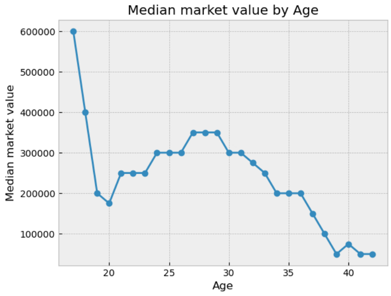
I searched for a correlation independently between a player’s height, weight, and BMI against the number of goals scored. I used the Pearson correlation coefficient to find out that every one of the three features has a very weak correlation coefficient with the number of goals scored, which means that neither Weight, Height nor BMI can explain the number of goals scored. I then retried using data only from players that have as general position ‘Attacker’, since they are the most concerned in scoring goals, and the relation was indeed higher but overall, still stayed very weak.

2.4 Visualisation

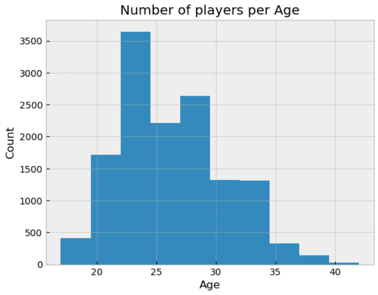
To find out if age is related to market value, I used a scatter plot first, which was hard to read information from since a lot of data-points were compressed together, so I also plotted the median market value against age on a line chart to have a better understanding of the distribution. By visualising the relation in two different ways we can see the main patterns. I also checked the distribution of Age using a histogram, from which we can conclude that most players are between 22 and 29 years old.



Graph 1. Market values by Age. (Source: NAC notebook)

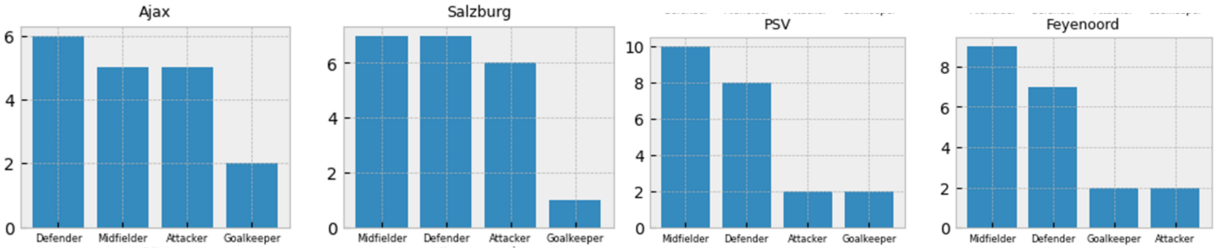


Graph 2. Median market value by Age. (Source: NAC notebook)



Graph 3. Number of players per Age. (Source: NAC notebook)

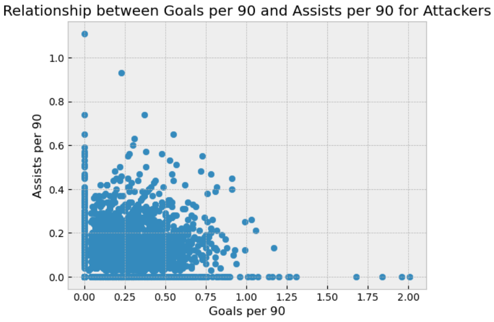
When looking at the distribution of players’ general positions across 4 different teams from the Netherlands using bar plots, we can see Ajax and Salzburg having an even and natural distribution of positions in their team, where it natural to have less goalkeepers because there is only one on the field, but important to have more other general positions. PSV and Feyenoord both seem to have a lack of Attacker in their team, which could be a problem when it comes to making changes.



Graph 4. Distribution of general positions across four different teams. (Source: NAC notebook)

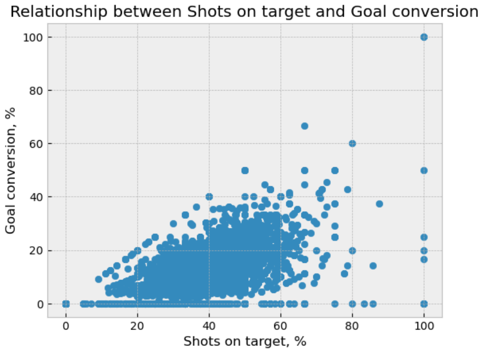
Using the Pearson correlation coefficient in combination with scatter plots I could find some additional interesting relations between our features:

- The relationship between goals and assists for attackers is very weak / almost not existent. This is because there are different kind of focusses the coach is giving his attackers: some attackers are specialised in assisting, some in scoring.
(Correlation coefficient: 0.05)



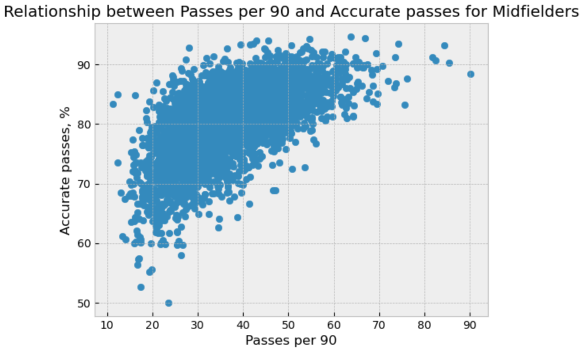
Graph 5, Relationship between goals and assists for attackers. (Source: NAC notebook)

- The relationship between shots on target rate and goal conversion rate for attackers is moderate.
(Correlation coefficient: 0.525)



Graph 6, Relationship between shots on target rate and goal conversion rate for attackers. (Source: NAC notebook)

- Lastly, the relationship between passes per 90 and accurate passes for midfielders is strong. In general, the more passes per 90 a midfielder makes, the more accurate those passes seem to be.
(Correlation coefficient: 0.629)



Graph 7, Relationship between passes per 90 and accurate passes rate for midfielders. (Source: NAC notebook)

2.5 Conclusion on EDA

Performing exploratory data analysis allowed us to understand the features of our dataset and make relations between them. EDA is a powerful tool for NAC since it helps to identify the players which have characteristics and performances that could be optimal for their team, it also helps to find under-valued players that have a strong growth potential. It is also essential know your data before beginning to work on machine learning models as it will influence how we select our features and our model choice.

3 Machine Learning

3.1 Data preparation for Machine Learning

Now that we have more insight on our dataset, I will remove the column we don't need (e.g. the column 'Player') and edit some columns to reduce dimensions (e.g. keeping only the year for column 'Contract expires'). It is important to mention that even though the birth country and passport country could make our prediction model better, I decided to delete those two features entirely from our dataset since it would be unethical to keep them as a player should not be worth less or more depending on his nationality.
We now have 7 categorical columns in our dataset, but most machine learning models work only with numerical data, so I will encode those categorical features into numerical data. Our new dataset contains 13753 different players with 183 features. One last important step before start training our models is to create a categorical target since we will be using linear models and but also non-linear ones. I thus created 5 market value bins: very cheap, cheap, normal, expensive, and very expensive.

3.2 Model evaluation

After trying out different models and comparing them, I decided to continue with two different classification models; XGBClassifier, which uses optimized boosting techniques to create an ensemble of decision trees, and Logistic Regression which is a classification model that uses linear combinations of features to predict the probability of a categorical outcome. Both models had the best scores when comparing all models.

To evaluate the models, I used multiple performance metrics: precision score, recall score and f1-score/accuracy score, which are used for evaluating classification models. I displayed those scores in a classification report where we can also observe those scores for each predicted bin. In addition to that, I plotted a confusion matrix to make it easier to understand our results. Finally, I used cross validation with 20 folds, to print out the mean cross validation score, which for the Logistic Regression model was 0.57 and for the XGBClassifier model 0.55.

3.3 Model improvement

To try improving both models I used hyperparameter tuning, which is a technique to find the optimal parameters resulting to the best score. More specifically, I used grid search on the Logistic Regression model to find the optimal 'inverse of regularization strength' and best 'solver'. For XGBClassifier I used grid search to find best parameters such as 'number of estimators', 'learning rate' and type of 'booster'.

Results:

- Logistic Regression:
  Mean cross validation score before hyperparameter tuning: 0.5706
  After hyperparameter tuning: 0.571

- XGBClassifier:
  Mean cross validation score before hyperparameter tuning: 0.5536
  After hyperparameter tuning: 0.5626

Hyperparameter tuning didn't impact the results of the model by much. My final best performing model stays thus Logistic Regression.

4 Ethical Considerations

4.1 Ethics

There are three elements vital for an ethical organisational capacity to NAC.
The first element is being an ethical company. To be considered ethical, a company should have respect for others, communicate openly, be responsible, transparent, and environment friendly. The parties responsible for this within NAC are the CEO and executives of the football club such as managers, HR-department, investors, and board of directors, who should constantly be searching to improve the ethicality of the club.
Secondly, the club should use ethical process and tools such as yearly ethics training programs that would inform all actors of the club about ethics and laws, or an ethic hotline as a way for employees to anonymously report suspected unethical behaviour. NAC could create trade unions for managing these tools or could collaborate with outside organisations specialised in ethical process.
The third element is to have ethical people in and around your football club. The first concerned are the employees and players from the club, they must adopt ethical and professional behaviours towards their customer, outside organisations with whom they are collaborating with, the environment, and their supporters. It also works the other way around, for NAC to be a professional club not only the inside of the club should be ethical but also the outside, e.g. racist and violent supporters would ruin the reputation of NAC and eventually the performance of their players. The key is continuously trying to create an ethical atmosphere around your football club until no critique is possible anymore, which indicates that the three elements are being respected.

On their official website, NAC claims to be aware of the social role they play in the city of Breda and take this seriously. They indeed currently have 8 different social non-lucrative ethical projects that contribute to the needs of the city. They have for example the 'Niemand buitenspel' project, which consists of offering the opportunity to gain work experience in their stadium to persons with disabilities or without starting qualifications, since it is often hard for them to find a job. Another ethical project is 'Playing for success Breda', which is a program for young children who are not preforming optimally at school for social-emotional reasons, that leads to more self-confidence and motivation of those children.

Ethical decision-making within this project has been done mainly through Right-based ethics. Indeed, the data we collected is publicly available, but more personal data about the players, like e.g. medical data has been kept private, thereby considering the GDPR rules about personal data. During the process of machine learning, I removed two features from our dataset that where unethical to consider in our prediction model. Predicting market value based on the birth country and passport country of a player be discriminating, as players should not be worth less or more depending on their nationality.

4.2 Recommendation to improve ethical guidelines.

To improve their ethical guidelines within the current project, I would recommend NAC, if that's not already the case, to allow their players to see their performance data more often. I believe it is important for a player to visually track his performances whenever he wants to develop a more personalised career. Managers will often present the performances of their players grouped, but only occasionally for a specific player, which can give an effect to the players of not having control over their performances. Creating an app for example where each player of the team could look at their stats freely whenever he wants, and where the stats are clear and visualized, would be a great tool not only in an ethical way, but also to push training further.

5 Recommendations

During this project we worked with data from 2022-2023, which gave us an insight on how players are performing this year (assuming this is the most updated data available), but it didn't show the evolution of those players. In most projects where we want to predict targets, we use historical data to track evolution, and to make our predictions follow that evolution. If we would have access to data the same data but yearly from 2015 to 2023 for example, we could perform deeper analysis e.g. we could select a specific goalkeeper and see how his saving rate evolves on an 8 year period, and better we could make better prediction of the market value since evolution would be an important contributor.

References:

Sean Peak. (2023, October 23). *A Culture of Ethical Behavior Is Essential to Business Success*. Business News Daily. https://www.businessnewsdaily.com/9424-business-ethical-behavior.html

Jake Wojtowicz. *Fans, Identity and Punishment*. Philarchive. https://philarchive.org/archive/WOJFIA

(2024). Syntrio. https://www.syntrio.com

GDPR Summary. (2020, December 6). *GDPR Summary – An overview of the General Data Protection Act.* https://www.gdprsummary.com/gdpr-summary/

Ethical guidelines for statistical practice. ASA. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice

Project Juptyer Notebook. https://github.com/BredaUniversityADSAI/2023-24b-fai1-adsai-LouieDaans232345/blob/main/final%20IPYNB%20draft%20/Final_Deliverable_Y1BlockB.ipynb

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