Implementing a data-driven solution for a modern problem

An insight into the analysis and application of machine learning algorithms for solving a business problem.

DISCOVER YOUR WORLD

Academy/Department

Saturday 01 January 2022

Breda University
OF APPLIED SCIENCES

Implementing a data-driven solution for a modern problem

An insight into the analysis and application of machine learning algorithms for solving a business problem.

Daria E. Vlăduțu

Department of Data Science and Artificial Intelligence, Breda University of Applied Sciences

Professional Writing

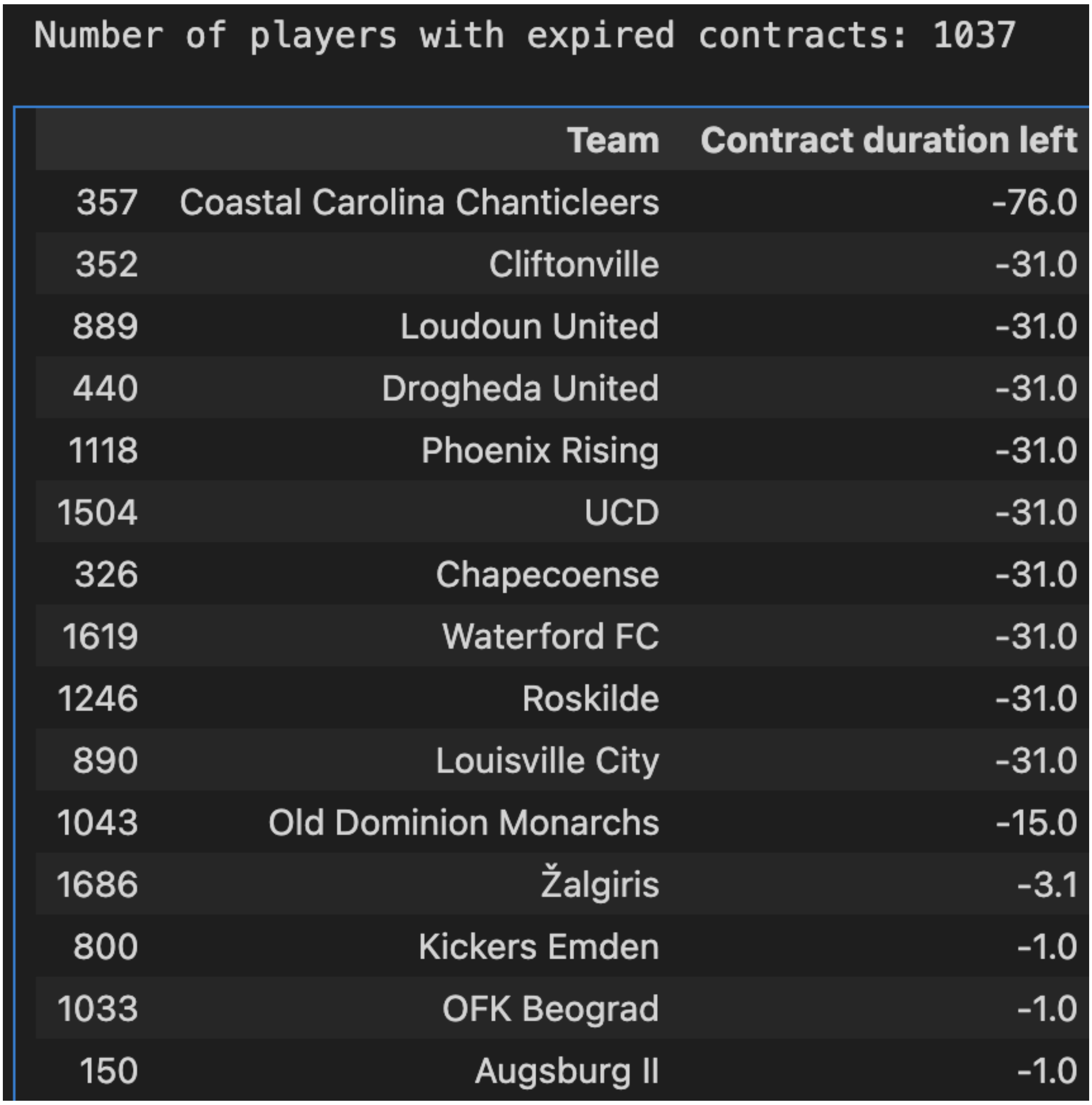
Arash Sadeghzadeh

January 26, 2024

Index

1 Introduction

Founded in 1912, NAC Breda stands as a staple in Dutch football, boasting a rich history primarily spent in the Eredivisie, the nation's top-tier football league. NAC Breda wants to improve its player recruiting approach by utilising data science. The proposed issue to solve is the acquisition of a player in the defender position at the lowest possible market value, using a configurable data-driven tool since the renowned football team will lose 1037 players by the expiry of the contract at the end of the season (information resulting from data analysis). With the help of Machine Learning algorithms, NAC Breda will be able to make more informed decisions throughout the hiring process and shape the club's future by focusing on data-driven excellence.



```
Number of players with expired contracts: 1037
```

	Team	Contract duration left
357	Coastal Carolina Chanticleers	-76.0
352	Cliftonville	-31.0
889	Loudoun United	-31.0
440	Drogheda United	-31.0
1118	Phoenix Rising	-31.0
1504	UCD	-31.0
326	Chapecoense	-31.0
1619	Waterford FC	-31.0
1246	Roskilde	-31.0
890	Louisville City	-31.0
1043	Old Dominion Monarchs	-15.0
1686	Žalgiris	-3.1
800	Kickers Emden	-1.0
1033	OFK Beograd	-1.0
150	Augsburg II	-1.0

Figure 1. Output of function for finding the number of players with expired contracts

The approach chosen for solving this problem is by creating multiple algorithms that aim to address these challenges by employing a Gradient Boosting Regressor model, optimizing its performance, and providing actionable insights for NAC's player acquisition strategy. Through this, the algorithm aims to contribute to NAC's competitive edge in the Dutch First Division by facilitating informed and ethical decision-making in player recruitment.

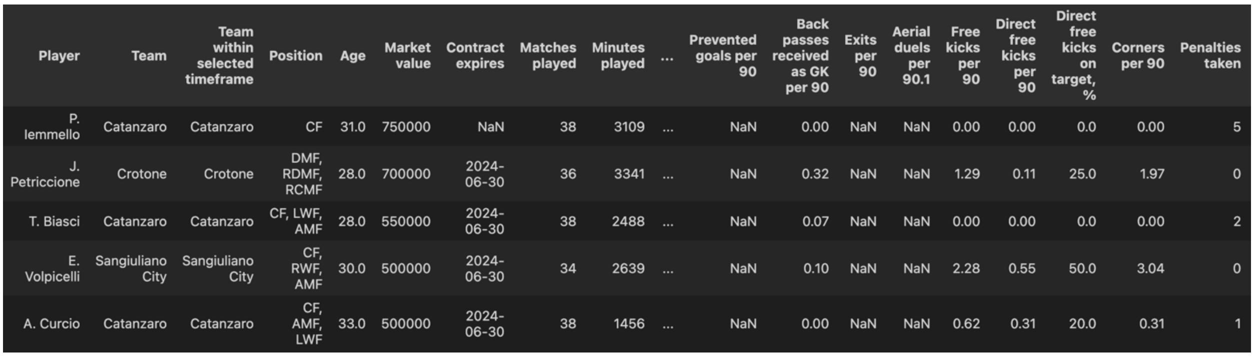
2 Exploratory Data Analysis

2.1 Overview of the dataset

The provided dataset originally contained 16535 rows and had 114 features. After preparing the data for analysis and applying the necessary methods, the new dataset contained 10443 and 114 features. The column named ‘Column expires’ was removed due to concerns about its usefulness in devising a solution for the presented business problem. However, a new feature called ‘Position category’ was added.

Initially, the data was presented in an Excel format which contained information about 45 countries’ football first division leagues. After iterating through all the files, they were appended to create a new data frame with the information that was used for further analysis. The resulting data frame was more suitable and concise, allowing for an easier exploration of the player data and improved visualisations which ultimately led to better conclusions.

Additionally, the NAC dataset contains 106 numerical columns and 9 categorical columns. Most of the numerical columns contain integer variables and float variables. However, the categorical columns mostly contain data that is an object.



Player	Team	Team within selected timeframe	Position	Age	Market value	Contract expires	Matches played	Minutes played	...	Prevented goals per 90	Back passes received as GK per 90	Exits per 90	Aerial duels per 90.1	Free kicks per 90	Direct free kicks per 90	Direct free kicks on target, %	Corners per 90	Penalties taken
P. Iemmello	Catanzaro	Catanzaro	CF	31.0	750000	NaN	38	3109	...	NaN	0.00	NaN	NaN	0.00	0.00	0.0	0.00	5
J. Petriccione	Crotone	Crotone	DMF, RDMF, RCMF	28.0	700000	2024-06-30	36	3341	...	NaN	0.32	NaN	NaN	1.29	0.11	25.0	1.97	0
T. Biasci	Catanzaro	Catanzaro	CF, LWF, AMF	28.0	550000	2024-06-30	38	2488	...	NaN	0.07	NaN	NaN	0.00	0.00	0.0	0.00	2
E. Volpicelli	Sangiuliano City	Sangiuliano City	CF, RWF, AMF	30.0	500000	2024-06-30	34	2639	...	NaN	0.10	NaN	NaN	2.28	0.55	50.0	3.04	0
A. Curcio	Catanzaro	Catanzaro	CF, AMF, LWF	33.0	500000	2024-06-30	38	1456	...	NaN	0.00	NaN	NaN	0.62	0.31	20.0	0.31	1

Figure 2. Head of NAC dataset

2.2 Processing of data

Some of the features used in the processing step were chosen based on an algorithm which performs feature selection using a Random Forest Regressor to identify the most important features for predicting the 'Successful defensive actions per 90' target variable. The target variable was chosen based on the subjective importance of features for a defender position.

2.2.1 Missing values

Missing values were identified for each feature column using the NumPy library which were then visualised using a bar chart for a better understanding. The following techniques were used to handle the missing values in the dataset:

- Rows where 'Passport country', 'Birth country', 'Age', or 'Position' have missing values were dropped;
- Missing values in the 'Foot' column were filled with 'No preference’;

- Missing values in the 'Team' column were filled with 'Free Agent';
- Missing values in the 'Position' column were filled with 'Unknown';
- The remaining missing values were filled with 0.

2.2.2 Outliers

Outliers were dealt with by applying a function which calculates the inter-quartile range for a specific column, calculates the outlier bounds, and filters the values in the column based on whether they fall within the calculated lower and upper bounds.

The 'Contract expires' column was removed because the algorithms use player statistics that are relevant for actual in-game performance. Therefore, it was deemed useless for the task at hand.

Moreover, the data was restricted to players who cost at most 300000 Euros. This decision was fuelled by research into the maximum value NAC spent on player acquisition in the last season (2022, Transfermarkt).

2.2.3 Normalization and Standardization

The data was normalized using the standard scaler/ z-score technique, and it was standardized using the robust scaling method. Categorical variables were one-hot encoded by creating new binary columns for each unique value (using Pandas' 'get_dummies' function), and by using a label encoder (using Scikit-learn).

A new data frame was created (entitled 'clean_NAC_data') where all the changes were saved, which was used further on for the implementation of the machine learning algorithms.

2.3 Data Statistics

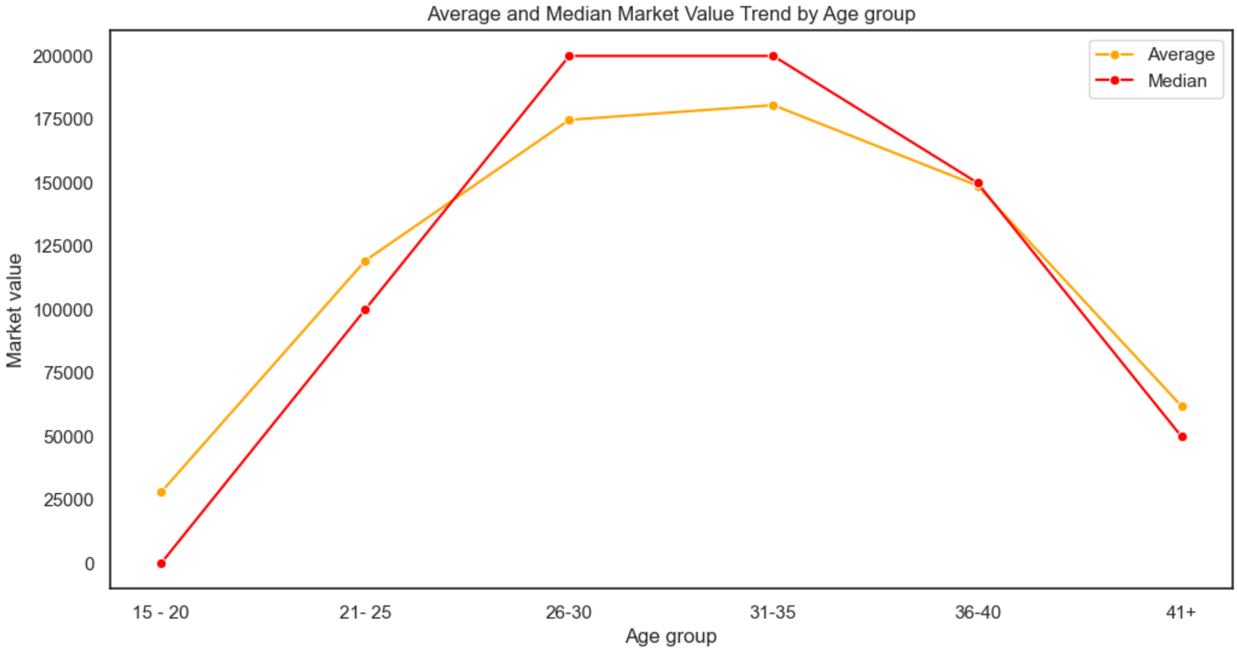


Figure 3. Line plot of average and median market value by age group

The median is higher than the mean in the represented line plot, suggesting a distribution where a few higher values (right skewed) are pulling the central tendency measure (median) towards the higher end of the data, which indicates the presence of a few unusually high values (outliers).

Skewed data may affect the performance of certain statistical models that assume normality, such as a Linear Regression model. Transformations such as normalization of data or standardization, or model adjustments may be necessary.

This insight can guide further analysis and decision-making related to player valuation and team management.

After analysing the data insights presented in the notebook, multiple conclusions can be drawn:

- The players' ages predominantly lie in the threshold of 20-25, suggesting that younger players are preferred. This is important to know when drafting an algorithm to aid in new player selection.

- The most predominant position category throughout the teams is defenders. This is the main reason why I choose to focus my algorithms on said position category.

- Most players' birth country for this season's data is Italy. This could serve as the indicator of which country's players to be researched more when it comes to player acquisition.

- Most duels and aerial duels (in percentages) are won by goalkeepers.

2.4 Visualisation techniques

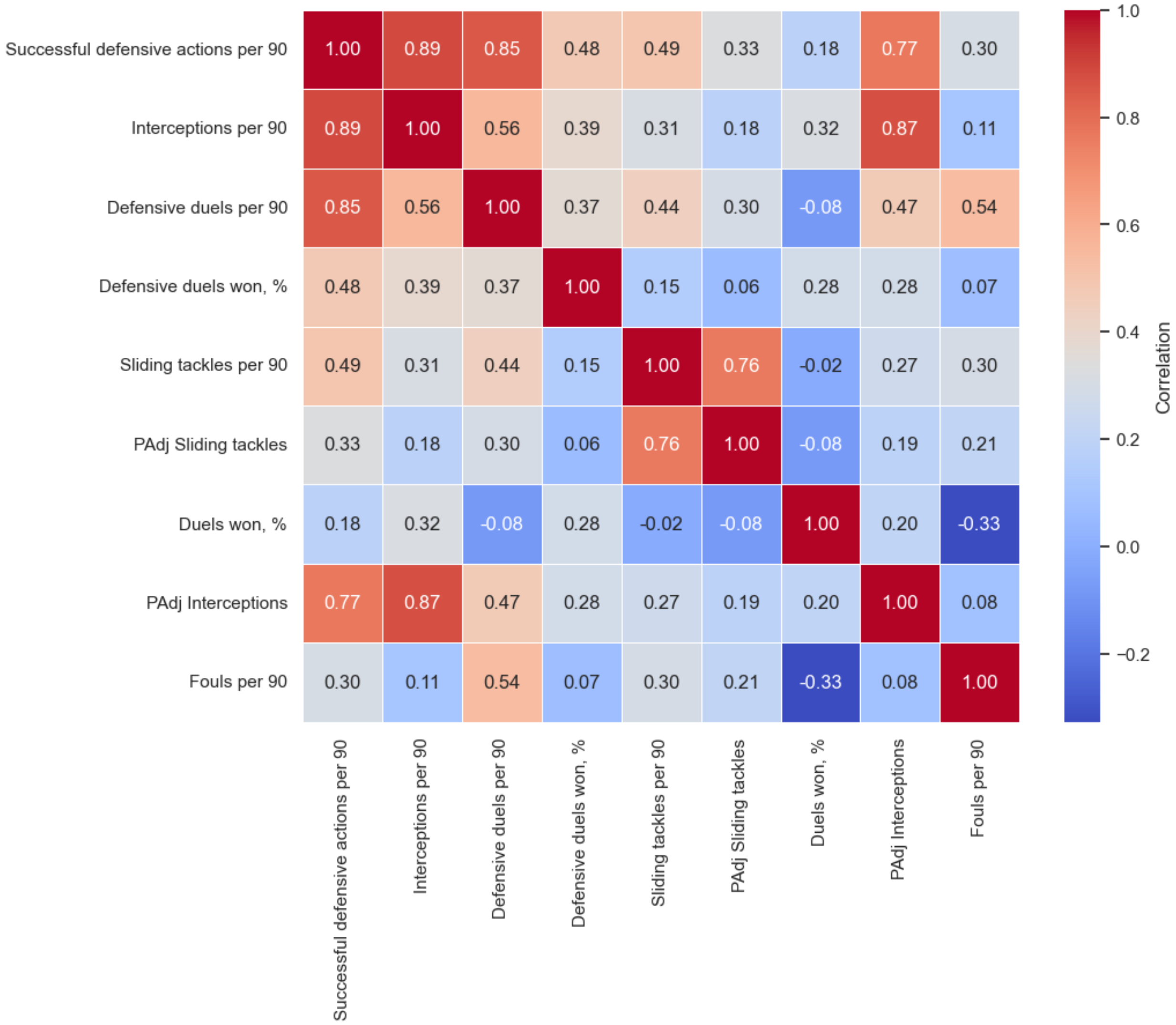
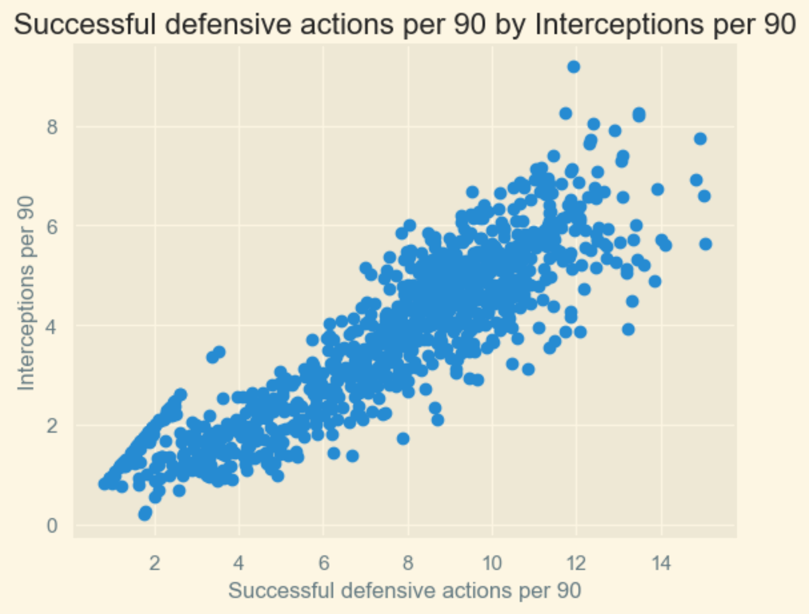
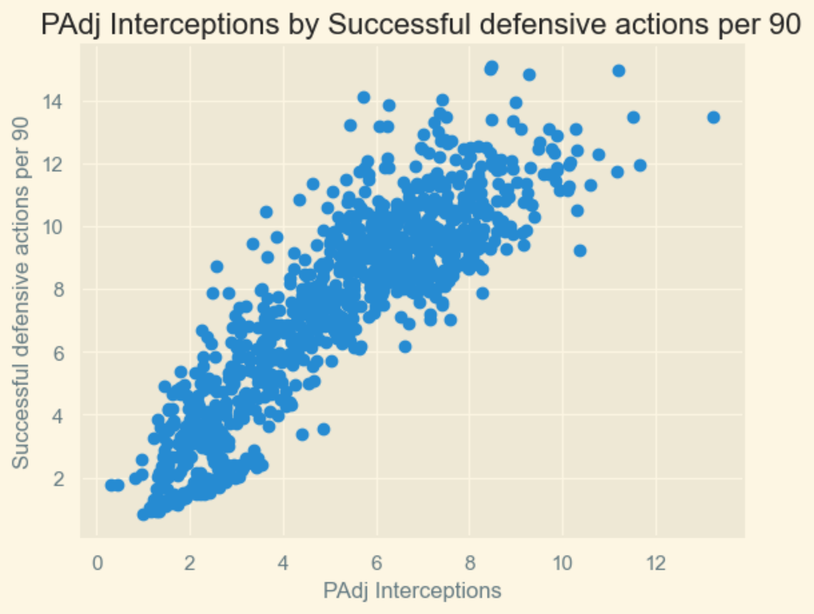


Figure 4. Correlation heatmap for the defender features

The presented heatmap of a correlation matrix indicated there is a high correlation between 'Successful defensive actions per 90' and 'Interceptions per 90', and between 'Successful defensive actions per 90' and 'Defensive duels per 90', indicating a strong linear relationship between these variables. This affected my choice of the Linear Regression model's variables.


Successful defensive actions per 90 by Interceptions per 90
Interceptions per 90
0
2
4
6
8
2
4
6
8
10
12
14
Successful defensive actions per 90


PAdj Interceptions by Successful defensive actions per 90
Successful defensive actions per 90
2
4
6
8
10
12
14
0
2
4
6
8
10
12
PAdj Interceptions

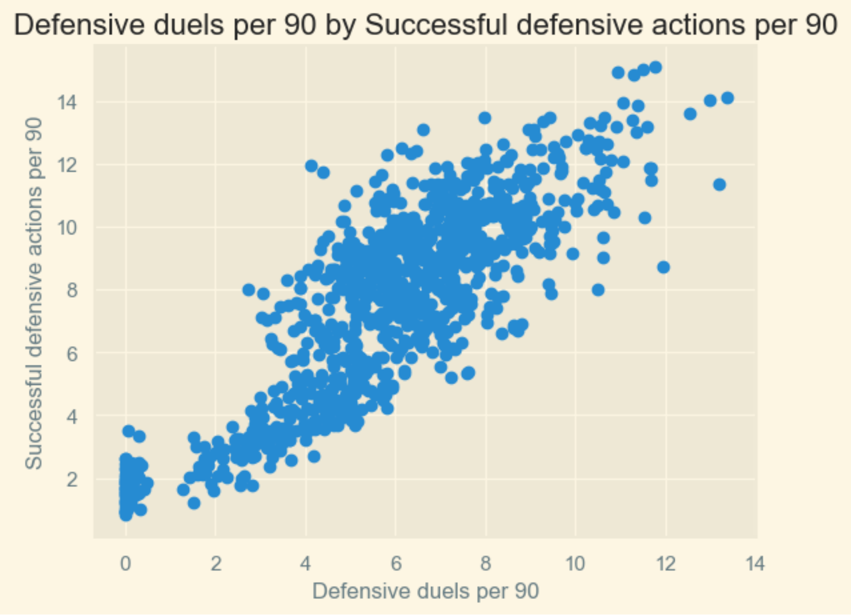


Figure 5, 6, 7. Scatter plots for the defender features with the highest correlation coefficient in relation to successful defensive actions per 90

These visualisations cemented the decision for the target variable for the Linear Regression model because there is a clear linear relationship between the plotted features and '*Successful defensive actions per 90'.*

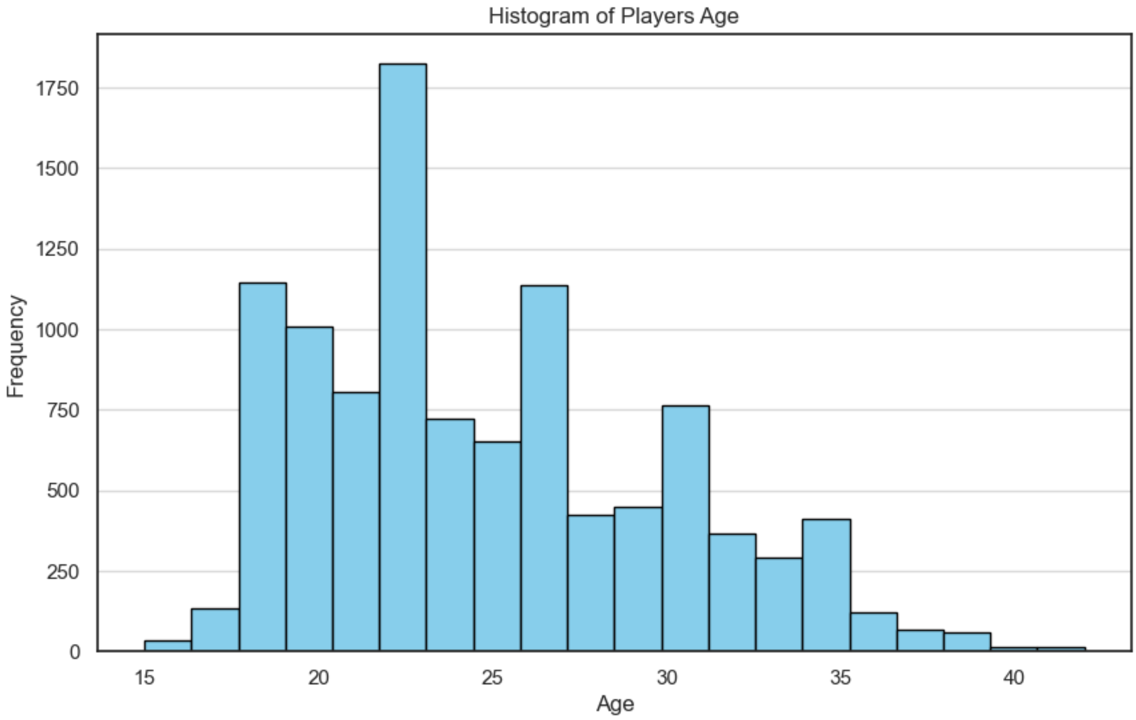


Figure 8. Histogram of players' age

The histogram is right-skewed, indicating that the majority of players are relatively younger, and there are a few veteran players or outliers with higher ages. This means that younger players are preferred when it comes to player acquisition. This influenced subsequent analysis because it underlined the importance of concentrating the player acquisition strategy on a specific age group.

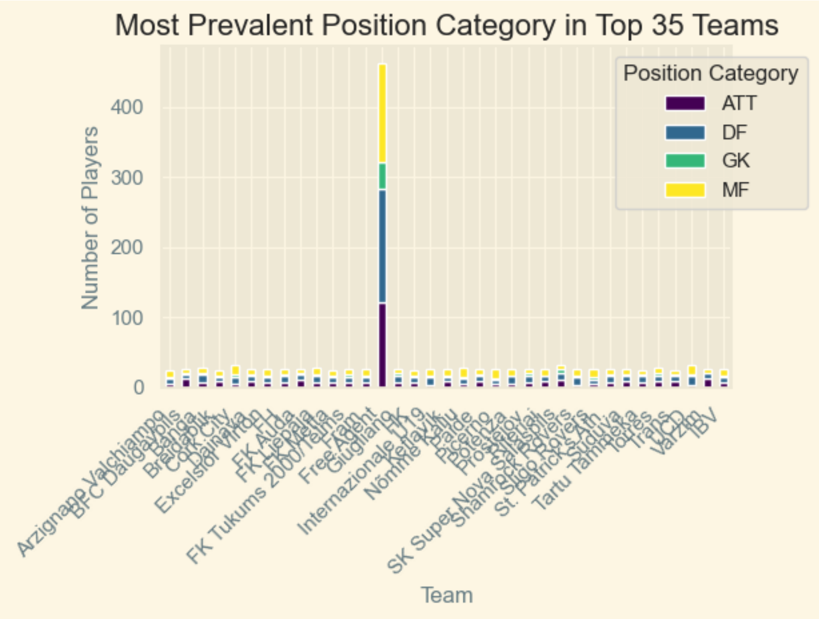


Figure 9. Stacked bar chart of the most prevalent position categories in the top 35 teams

Based on this visualisation where it is clear that the defender position is the most prevalent position category in the top 35 teams, the data frame used for training the models was reduced to include only said category.

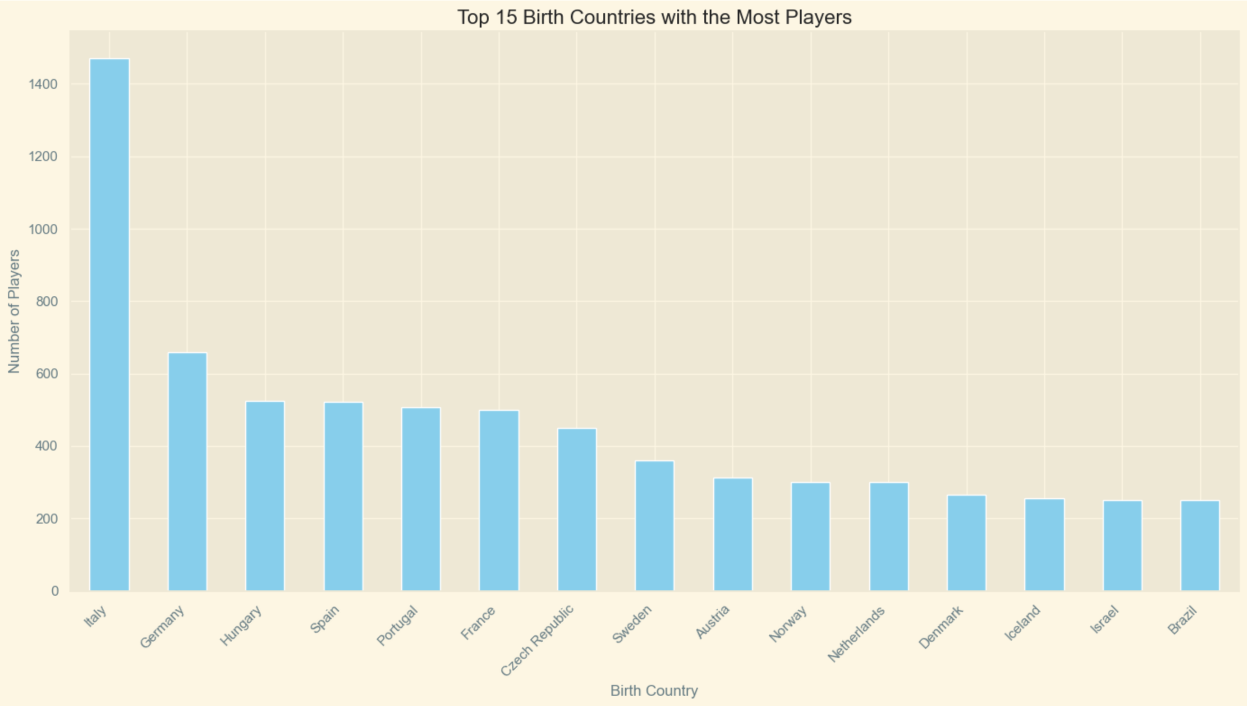


Figure 10. Bar chart of the top 15 birth countries with the most players

Italy is the birth country of most players included in the dataset for this season, which implies that NAC should focus their search efforts for new talent in said country.

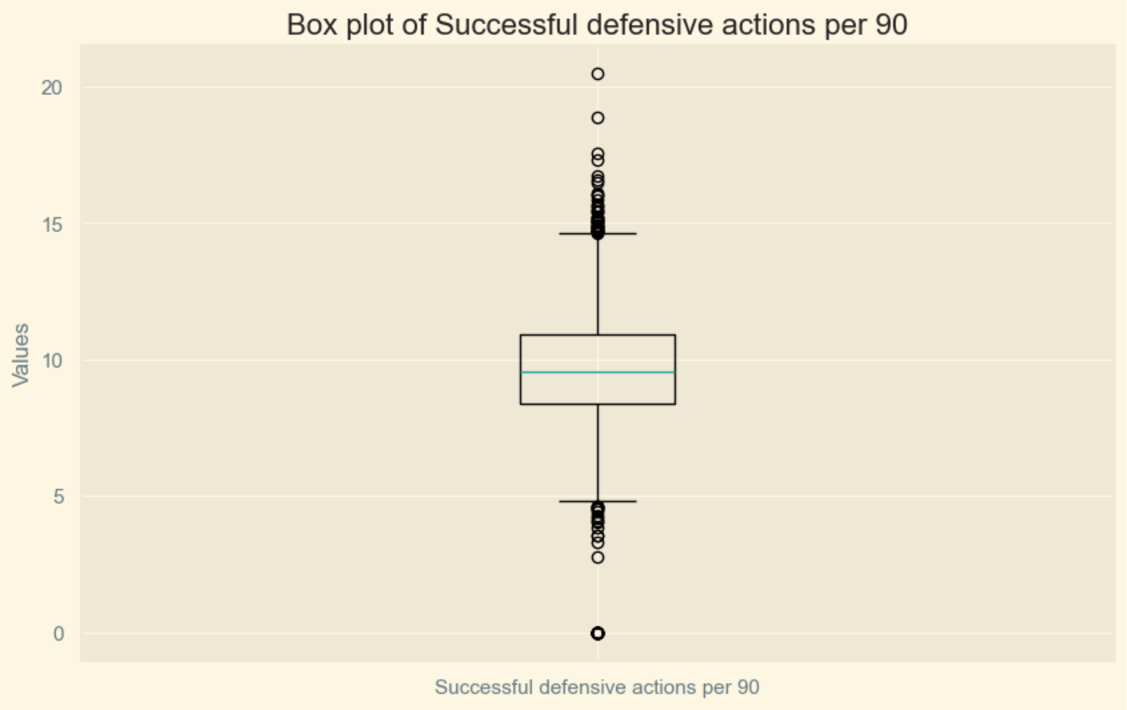


Figure 7. Box plot of successful defensive actions per 90

The visualisation presented facilitated the identification of outliers for the target variable, which is very important not only for the model training but also for the interpretations of the model performance metrics.

2.5 Examining the relationship between variables

After calculating the correlation coefficient for the most important features of a defender, calculated using a Random Forest Regressor algorithm, these are the conclusions drawn: there is a high correlation between 'Successful defensive actions per 90' and 'Interceptions per 90' (*0.89*), between 'Successful defensive actions per 90' and 'Defensive duels per 90' (*0.85*), and between 'Successful defensive actions per 90' and ‘PAdj Interceptions’ (*0.77*), indicating a strong and positive linear relationship between these variables. Therefore, this information is useful when it comes to inputting the features for training the Logistic Regression model.

On the contrary, there is a small correlation between 'Successful defensive actions per 90’ and ‘Duels won’ (in percentages) (*0.18*), which is counterintuitive, and it suggests a weak or negligible linear relationship between these two variables.

However, as correlation does not mean causation, the correlation coefficient does not communicate all the information necessary for the models to perform well.

2.6 Summary of key findings from the EDA

Several significant trends and patterns were revealed by the study of the football dataset. Notably, a right-skewed age distribution highlighted a preference for younger players, and the importance of this position category was shown by the high percentage of defenders in the sample. The investigation also revealed trends in the birth countries of the players, with Italy emerging as an important contributor. Potential theories derived from the analysis include the significance of age groups in team planning and the impact of a birth nation on player acquisition tactics. Furthermore, the existence of outliers in defensive plays that work emphasises how crucial it is to comprehend their characteristics to make strategic decisions in football management. All things considered, these revelations will drive the course of further investigations, directing feature selection and impacting model selections for improved relevance and accuracy in the field of football.

3 Machine Learning

3.1 Method

3.1.1 Chosen model

The machine learning model that would be best suited for the business problem presented is the Gradient Boosting Regressor model. The target variable is ‘Successful defensive actions per 90’, and the features used are: 'Interceptions per 90', 'Defensive duels per 90', 'Defensive duels won, %', 'Sliding tackles per 90', 'PAdj Sliding tackles', 'Duels won, %', 'PAdj Interceptions', 'Fouls per 90'.

The model was chosen on account of the task at hand being a regression task. The algorithm works by building a predictive model in the form of an ensemble of weak learners, typically decision trees, and sequentially adds them to correct the errors made by the existing ensemble, working well for regression tasks. Additionally, the Gradient Boosting Regressor is known for its flexibility and high predictive performance, making it an ideal candidate for aiding in the acquisition of new players.

3.2 Model evaluation

The metrics used for evaluation of the GBR model were: Mean Squared Error, used due to its sensitivity to outliers therefore emphasizing larger errors; Root Mean Squared Error, which is essentially identical to the MSE, the only difference being the easier interpretability since it is in the same unit as the target variable; and R squared, which measures the variance in the dependent variable that is predictable from the independent variables, making this metric particularly useful for understanding the proportion of variability in the data captured by the model. With an initial MSE of approximately 0.044, an RMSE of approximately 0.21, and an R squared value of approximately 0.98, it is clear the model is performing very well on the given data. In summary, the low MSE and RMSE values indicate good predictive accuracy, and the high R-squared value suggests that the Gradient Boosting model is capturing a large proportion of the variance in the dependent variable.

3.2.1 Cross-validation techniques

K-fold cross-validation was used for assessing the performance of the model (with 5, 7, and 10 folds), because it helps massively in identifying under- or overfitting. With an accuracy for each n-fold mentioned of around 8%, the model's performance in correctly classifying instances is relatively low.

On the other hand, the MSE, RMSE, and R-squared values are indicative of good predictive performance for regression tasks. The model appears to be highly accurate in predicting continuous numerical values, which is what is desired.

3.3 Model improvement

For tuning, the following hyperparameters were used to improve the Gradient Boosting Regression:

- 'n_estimators': [50, 100, 150],
- 'learning_rate': [0.01, 0.1, 0.2],
- 'max_depth': [3, 4, 5],
- 'min_samples_split': [2, 5, 10],
- 'min_samples_leaf': [1, 2, 4].

The parameters were chosen based on subjective perception of importance for the task at hand and randomly assigned values.

Grid Search was the technique used for hyperparameter optimization. There were no major challenges encountered when applying the technique.

Results of the search:

- Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 150}

- New Mean Squared Error: 0.0373677708033465
- Old Mean Squared Error: 0.044887622555627935

- New Root Mean Squared Error: 0.19330745149462425
- Old Root Mean Squared Error: 0.21186699260533232

- New R squared: 0.99114231921877
- Old R squared: 0.9893598086511898

- Accuracy for each fold for 5-fold cross-validation: 0.08828703894365451
- Accuracy for each fold for 7-fold cross-validation: 0.0878250671558453
- Accuracy for each fold for 10-fold cross-validation: 0.08771981602101833

After inputting the best hyperparameters into the Gradient Boosting Regressor model, and evaluating it, there has been a decrease in the mse value by approx. 0.01, a decrease in the rmse by approx. 0.02 and an increase in the R-squared value by approx. 0.01.

4 Ethical Considerations

Three key components for ethical organisational capacity—transparency, accountability, and fairness—were found while analysing the ethical issues in the project and its applicability to NAC. Fairness is about treating everyone equally, accountability is about assigning clear responsibilities, and transparency is about open communication and information exchange. Fairness is a shared duty among all team members in NAC, accountability is the responsibility of management and team leaders, and openness is the responsibility of the communication and public relations team. The study's conclusions show that NAC has incorporated ethical principles by upholding open lines of communication, delegating tasks, and encouraging fairness among its members.

A framework was established, paying particular attention to GDPR and Ethical Guidelines for Statistical Practice, to ensure ethical decision-making. To comply with GDPR rules, the project followed the guidelines of informed permission, data anonymization, and safe data processing. Maintaining data integrity, abstaining from deception, and offering correct interpretations were all done under ethical principles for statistical practice.

Despite these positive aspects, some ethical problems were identified within NAC. These include potential biases in player selection processes, data privacy concerns, and the need for more comprehensive ethical training for team members. To address these issues, recommendations for NAC include implementing diversity and inclusion initiatives, enhancing data protection measures, and providing continuous ethical training programs for all personnel involved in the project. By integrating these recommendations, NAC can strengthen its ethical guidelines, fostering a culture of transparency, accountability, and fairness in all aspects of its operations.

5 Recommendations

In order to study these theories more thoroughly, more data analysis might look at how players from different birth countries perform differently and how the success of a team is affected by particular age groups. These results highlight the need to select features carefully during model construction, paying particular attention to factors such as birth country, age, and position.

To improve the algorithm, NAC should prioritise variables that are special to defenders, including "Interceptions per 90," "Defensive duels per 90," and others, by utilising insights from the Gradient Boosting Regressor model. Prioritising the acquisition of younger players in the 20–25 age bracket is consistent with the trend in player preferences that has been noted. Increasing defender scouting efforts would help acquire exceptional players, especially because defenders are represented on 35 clubs. Acknowledging the prevalence of Italian players calls for a targeted scouting approach in Italy, maybe in conjunction with regional football organisations. In order to maintain equity, accountability, and transparency, staff members engaged in player recruiting must get ongoing ethical training that addresses possible biases and privacy issues. The execution of diversity and inclusion programmes and improved data security protocols will strengthen NAC's ethical base even further. Sustaining success in player acquisition while upholding moral standards will need constant observation and assessment in addition to flexibility in response to changing football dynamics.

6 References

Ashok, A. (2023, March 14). *NAC Breda 2022/23: How the Dutch club can now aim for promotion again using Thomas Tuchel-esque Tactics - Scout Report*. Total Football Analysis Magazine. https://totalfootballanalysis.com/team-analysis/nac-breda-2022-23-their-tactics-under-peter-hyballa-tactical-analysis-scout-report-tactics-analysis

Belyh, A. (2022, September 20). *Key elements of business ethics*. FounderJar. https://www.founderjar.com/key-elements-of-business-ethics/

NAC Breda. (n.d.). *De Club*. https://www.nac.nl/de-club

NAC Breda - profilul clubului. Transfermarkt. (n.d.). https://www.transfermarkt.ro/nac-breda/startseite/verein/132

Official Legal Text. General Data Protection Regulation (GDPR). (2022, September 27). https://gdpr-info.eu/

OpenAI. *ChatGPT*. Source text was written by me and then summarized. Prompt: ‘Summarize this text for me', (26-01-2024).

All figures presented are made by the author of this report using Python libraries such as Matplotlib and Seaborn, and no other external sources.

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