NAC Breda — Player Analytics & Market Value Modeling

Author: Mohamed Elshami

Course: Year 1, Block B — ADSAI

Date: **January 2024**

Executive Summary

This project consolidates all Block B Python work into a single, reproducible analysis focused on NAC Breda player analytics.
I cleaned and combined a large, multi-feature football dataset (16k+ rows, 100+ features), explored relationships between age, goals, xG/xA, contract duration, and market value, and implemented several baseline machine-learning models.

Key outcomes:
- A cleaned master dataset (`Combined_DATA.csv`) with numeric and categorical NA handling, type fixes, deduplication, and consistent schema.
- EDA insights (scatter, bar, hist, clustered scatter) describing patterns for NAC Breda and across the full dataset.
- Feature selection identified performance-relevant predictors for market value (e.g., Goals, xG, Assists, xA, Non-penalty goals).
- Modeling: Multiple baselines (Linear Regression, Random Forest, Gradient Boosting, XGBoost, SVM) and a full preprocessing pipeline. PCA slightly reduced MSE.
- Limitations: Market value is noisy, non-linear, and influenced by external factors; logistic regression is not well-suited to continuous targets; baselines show room for improvement.

1. Business Case & Objective

Goal: Support NAC Breda recruitment and valuation by understanding the factors associated with a player’s market value and on-pitch contribution.
Questions:
1. What relationships exist between Age and Market value?
2. How do goal-related features (Goals, xG, xA, Non-penalty goals) relate to value?
3. What team-level patterns (e.g., contract duration left, age groups) might inform retention or scouting?
Intended Use: Provide exploratory evidence and baseline predictive models to guide further analysis and dashboarding.

2. Data: Sources, Cleaning, and Management

Data sources included player performance and market valuation datasets.
Cleaning involved numeric/categorical NA imputation, type conversions, deduplication, and exporting a combined dataset (Combined_DATA.csv).

3. Exploratory Data Analysis

Key EDA highlights:
- Market value vs. age: Value peaks mid-20s with high variance.
- Height, weight, goals: Positive correlation with physical attributes.
- Top xG vs. Goals: Reveals finishing efficiency.
- Contract duration: Identifies renewal priorities.
- Matches by age: Mid-20s peak consistency.
- Player age distribution: Majority between 20–27 years old.
- Birth countries: Italy, Germany, and France dominate.
- Clustering: Segments based on Age & Market Value, Matches & Goals.

4. Feature Engineering & Selection

Top predictors for Market Value:
- Goals
- xG (Expected Goals)
- Assists
- xA (Expected Assists)
- Non-penalty goals

These features balance actual output and expected contribution for holistic evaluation.

5. Modeling

Models implemented:
- Linear Regression (best performer)
- Random Forest
- Gradient Boosting
- XGBoost
- SVM (classification baseline)

Results:
- Linear Regression MSE: 1.67e12
- Random Forest MSE: 1.91e12
- Gradient Boosting MSE: 1.78e12
- PCA + Linear Regression MSE: 1.62e12

Interpretation:

Linear Regression provided the best balance of accuracy and interpretability.

6. Interpretation & Discussion

EDA confirms logical football insights, but predicting market value is complex due to subjective and external influences.
Linear models outperform trees given the current feature set, but richer inputs (minutes, competition level) could improve results.

7. Ethical, Legal, and Operational Considerations

- Data fairness: Bias risk toward attacking roles.
- Privacy: Ensure GDPR compliance with player data.
- Transparency: Explainable models preferred.
- Deployment: Should complement, not replace, scouting decisions.

8. Limitations

- Target variable noise (market value subjectivity).
- Missing external factors (injuries, salary, competition level).
- Baseline hyperparameters not optimized.

9. Recommendations & Next Steps

1. Expand features: Minutes per 90, defensive metrics, progressive passes.
2. Test advanced models: Ridge, Lasso, SVR, LightGBM, CatBoost.
3. Apply SHAP for interpretability.
4. Build a Streamlit dashboard for NAC Breda scouts.