# Data-driven football scouting: *A role-based neural network approach*

Jacopo Savaia DA: Dr Fawada Qaiser DL: Dr Kathleen Kelm
CSCK700 Computer Science Capstone Project June 2025

## Introduction

Player recruitment is a high-risk and high-cost decision process for professional football clubs. Mid-table clubs in particular, must identify players who are likely to progress to higher-level leagues before their market value increases. Traditional scouting relies heavily on subjective judgement, while existing data-driven approaches often optimise for classification accuracy rather than practical scouting usefulness.

This project proposes a role-specific, ranking-oriented machine learning framework to identify players with a high probability of upward career movement, focusing on midfield roles. Rather than predicting exact career outcomes, the system prioritises producing high-quality shortlists, reflecting real scouting workflows.

## Research Questions

1. To what extent can a neural network trained on season-level performance data predict a football player's progression across league tiers?
2. Which technical performance indicators (such as passing accuracy, duels won, or dribbles per 90) most contribute to early identification of high-potential players within specific roles?
3. How can supervised learning models transform historical scouting data into actionable insights that support recruitment decision-making?

## Data & Problem Formulation

**Dataset:** Ligue 1 Seasons: 2017/18 - 2018/19 - 2019/20 - 2020/21 - 2021/21
**Inclusion criteria:** Age: 18–28 Minutes played ≥ 700 Statistics normalised per 90 minutes
Each row represents a player-season.
**Target Variable:** *Multiclass Label*: 2 = Player subsequently moved to a higher-tier league or team, 1 = Player moved horizontally to same level league or team, 0 = Player did not move upward or laterally
*Binary Label* focuses instead on finding the class 2 label rather than classify correctly the whole dataset.
**Role Segmentation:** Models are trained separately for: Attacking Midfielders (att_mid) - Defensive Midfielders (def_mid). This avoids diluting signals across fundamentally different playing responsibilities.

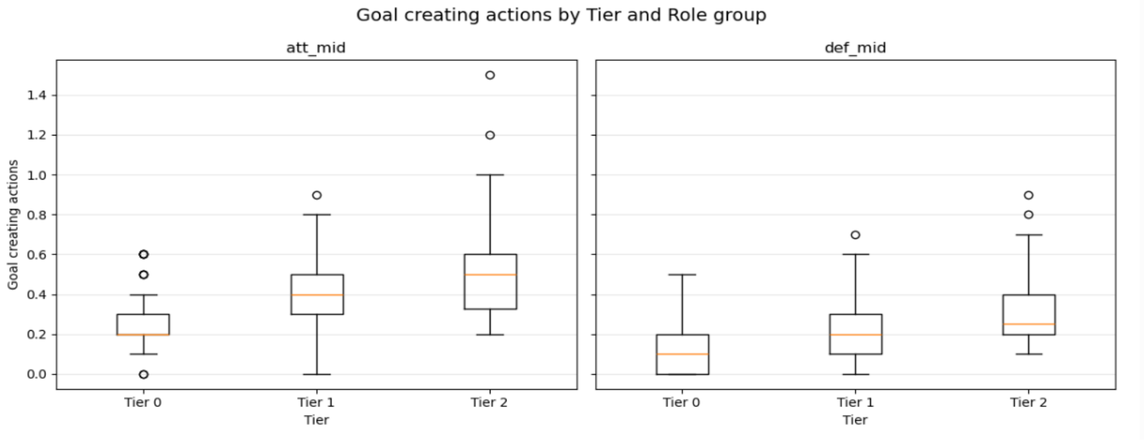
## Methodology

**Data Collection:**
Manual scraping from statshead.com
**Data Preprocessing:**

- Manual feature selection and feature engineering.
- Career progression labels were manually constructed using transfer outcomes following each season and averaged over a period of 3 years.
- The final dataset contains approximately 500 player–season observations.
- To account for fundamentally different tactical responsibilities, players were divided into:
  *Attacking midfielders*
  *Defensive midfielders*
- Separate models were trained per role group, avoiding signal dilution and improving interpretability.

**Model Development:**

- Players were labelled as upward movers (Tier 2 Label) if they subsequently transferred to a higher-tier league or team within a 3 years timeframe from the selected season, for lateral movement Tier 1 Label and downwards movemet Label 0.
- Multiclass mode were trained following this logic andl offered modest performances.
- A new binary class approach was implemented showing promising results.
- Binary classification models were trained to predict the probability of upward transfer. XGBoost was used as the primary model due to its strong performance on tabular data. Neural networks were also explored for comparison.
- Rather than selecting models using accuracy or F1 score, hyperparameters are tuned to maximise ranking quality, reflecting scouting priorities.
- Model outputs were interpreted as probabilities of upward transfer (p_tier2).




**Evaluation Metrics:**

- Model performance was evaluated using ranking-based metrics:
  *NDCG@20* to assess ranking quality of top candidates
  *Hit Rate@20* to measure shortlist recall
- Accuracy and Macro-F1 were reported for reference but not optimised
- Players were ranked by p_tier2 to generate scouting shortlists, rather than using hard class predictions

## Results

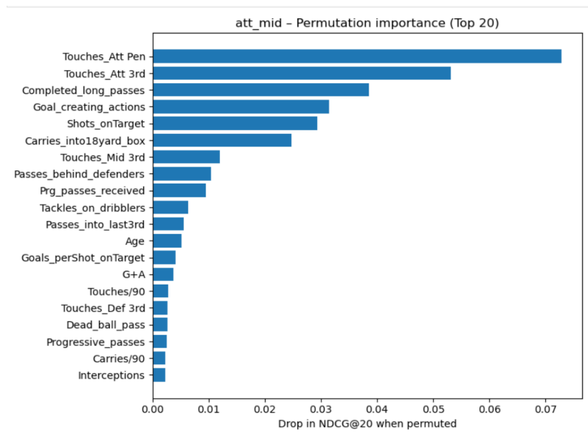
Binary XGBoost model performance table

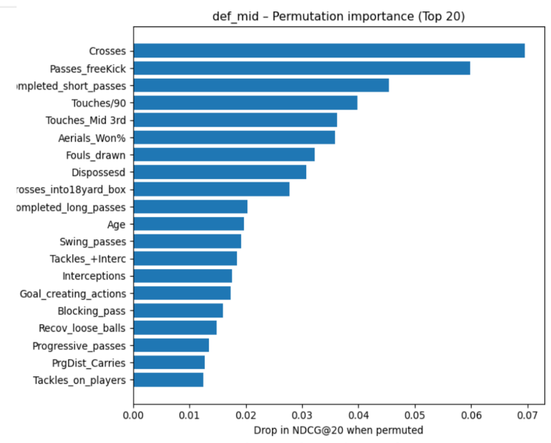
| | accuracy | macro_f1 | hit_rate@10 | ndcg@10 | hit_rate@20 | ndcg@20 |
|---|---|---|---|---|---|---|
| **def_mid** | 0.827586 | 0.764610 | 0.8 | 0.975265 | 0.45 | 0.971824 |
| **att_mid** | 0.854167 | 0.799881 | 0.7 | 0.897150 | 0.50 | 0.890137 |

Top 10 players suggestion scouting style table for attacking and defensive midfielders on test set.
The model will be finally tested with unseen data from 2024/25 season to get real scouting suggestions.

| | Player | Team | Role | role_group | true_is_tier2 | p_tier2 |
|---|---|---|---|---|---|---|
| 0 | Aleksandr Golovin | Monaco | AM | att_mid | 1 | 0.916521 |
| 1 | Romain Del Castillo | Rennes | LW | att_mid | 0 | 0.886558 |
| 2 | Armand Lauriente | Lorient | AM | att_mid | 1 | 0.863158 |
| 3 | Jonathan Ikone | Lille | RW | att_mid | 1 | 0.842699 |
| 4 | Thomas Lemar | Monaco | AM | att_mid | 1 | 0.798486 |
| 5 | Neymar | Paris S-G | AM | att_mid | 1 | 0.689580 |
| 6 | Kamaldeen Sulemana | Rennes | LW | att_mid | 0 | 0.685368 |
| 7 | Giovani Lo Celso | Paris S-G | AM | att_mid | 1 | 0.625676 |
| 8 | Lovro Majer | Rennes | AM | att_mid | 1 | 0.621724 |
| 9 | Luiz Araújo | Lille | RW | att_mid | 1 | 0.518157, |

| | Player | Team | Role | role_group | true_is_tier2 | p_tier2 |
|---|---|---|---|---|---|---|
| 0 | Fabinho | Monaco | CM | def_mid | 1 | 0.896039 |
| 1 | Marco Verratti | Paris S-G | DM | def_mid | 1 | 0.892985 |
| 2 | Khéphren Thuram | Nice | DM | def_mid | 1 | 0.890993 |
| 3 | Tanguy Ndombele | Lyon | CM | def_mid | 1 | 0.853494 |
| 4 | Azzedine Ounahi | Angers | CM | def_mid | 0 | 0.804726 |
| 5 | Adrien Tameze | Nice | DM | def_mid | 1 | 0.773882 |
| 6 | Ibrahim Sangaré | Toulouse | CM | def_mid | 1 | 0.762366 |
| 7 | Valentin Rongier | Nantes | CM | def_mid | 1 | 0.716900 |
| 8 | Yacine Adli | Bordeaux | CM | def_mid | 0 | 0.684195 |
| 9 | Youri Tielemans | Monaco | CM | def_mid | 1 | 0.556934) |




Feature importance was computed using permutation importance, measured as the drop in NDCG@20 after feature shuffling, to reflect the model's ranking objective.

## Conclusions

| Model | Role | Accuracy | Hit@20 | NDCG@20 |
|---|---|---|---|---|
| Binary XGBoost | att_mid | ~ 0.85 | 0.5 | 0.90 |
| Multiclass XGBoost | att_mid | 0.54 | 0.35 | 0.81 |
| MLP (Binary) | att_mid | 0.71 | 0.35 | 0.93 |
| Binary XGBoost | def_mid | 0.83 | 0.45 | 0.97 |
| Multiclass XGBoost | def_mid | 0.57 | 0.5 | 0.80 |
| MLP (Binary) | def_mid | 0.74 | 0.5 | 0.92 |

- A role-specific, ranking-based machine learning approach effectively identifies players with upward transfer potential.
- Binary XGBoost models, optimised using NDCG@20, outperform multiclass and neural network alternatives for scouting shortlists.
- Ranking players by predicted probability (p_tier2) aligns model outputs with real scouting workflows.
- Permutation importance reveals role-dependent progression signals, combining ball involvement, progression actions, and age.
- The framework demonstrates how interpretable ML models can support evidence-based football recruitment decisions.

References:
Lacan, S., 2023. DNN for scouting in football: Predicting player progression using neural networks. MSc Dissertation, University of Amsterdam. https://doi.org/10.48550/arXiv.2403.08835
McElfresh, D., Elenberg, E., Al-Shedivat, M. and Dubey, A., 2023. When Do Neural Nets Outperform Boosted Trees on Tabular Data? Proceedings of the NeurIPS 2023 Track on Datasets and Benchmarks. Available at: https://arxiv.org/pdf/2305.02997
Aparna Dhinakaran (2023).'Demystifyng NDCG' Towards Data Science. Jan 25. Available at: https://towardsdatascience.com/demystifying-ndcg-bee3be58cfe0/ (Accessed on: 22/12/2025)