Predicting Football Player Market Value

Position-Specific Machine Learning Pipeline

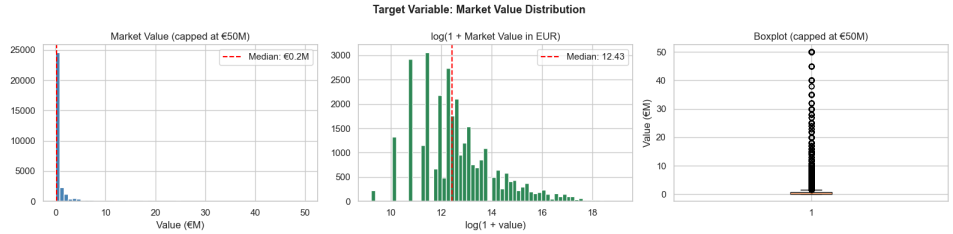
DAMA Hackathon 2026

MSc Data Science & Machine Learning

The Problem

Why is this hard?

- Transfer fees exceed **€200M** for top players
- Valuations are **subjective** — clubs, agents, media shape prices
- A striker's value depends on **different signals** than a goalkeeper's



Market value distribution — heavily right-skewed; log-transform applied

Our Goal

> Predict `market_value_in_eur` from **publicly available** performance statistics — **separately per position**

Data & Pipeline

7 Transfermarkt CSVs

Source	Size
Players	34,291
Appearances	1.7M rows
Valuations	449K rows
Clubs / Transfers	~85K

After cleaning: 30,718 labelled players

4 position groups:

GK · DEF · MID · ATT

10-step pipeline

```
Raw CSVs
  ↓ 01 Sanity checks
  ↓ 02 EDA
  ↓ 03 Clean & merge
  ↓ 04 Feature engineering
  ↓ 05 Preprocessing
  ↓ 06 Baseline models
  ↓ 07 Advanced models
  ↓ 08 SHAP & evaluation
  ↓ 09 Undervalued players
  ↓ 10 Final report
```

Feature Engineering

Shared base features (all positions)

- `age` + `age²` — non-linear career arc
- `log_minutes`, `log_appearances` — experience
- Rate stats: goals, assists, contributions per 90
- `log_highest_mv` — career peak reputation
- `league_mean_value` — target-encoded league prestige
- `transfer_count`, `log_total_fees`

Position-specific additions

GK · save percentage proxy, clean sheet rate, offensive contribution

DEF · defensive solidity, discipline score, defensive attack rate

MID · creative output per 90, attack-defense ratio

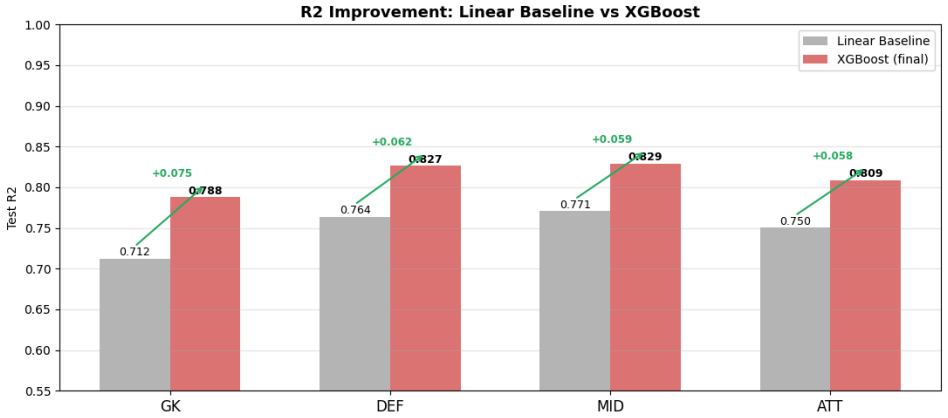
ATT · goals per appearance, minutes per goal, assist-to-goal ratio

Models & Results

What we trained

1. **Ridge** (linear baseline, RidgeCV)
2. **Random Forest** (RandomizedSearchCV)
3. **XGBoost** (hist method, tuned)
4. **MLP** (Adam, early stopping, 3 layers)

No data leakage: split → encode → scale



XGBoost gains +5–8 R² points over Ridge baselines

Test-set R² by position

Position	Ridge	RF	MLP	XGB
GK	0.712	0.781	0.775	**0.788**

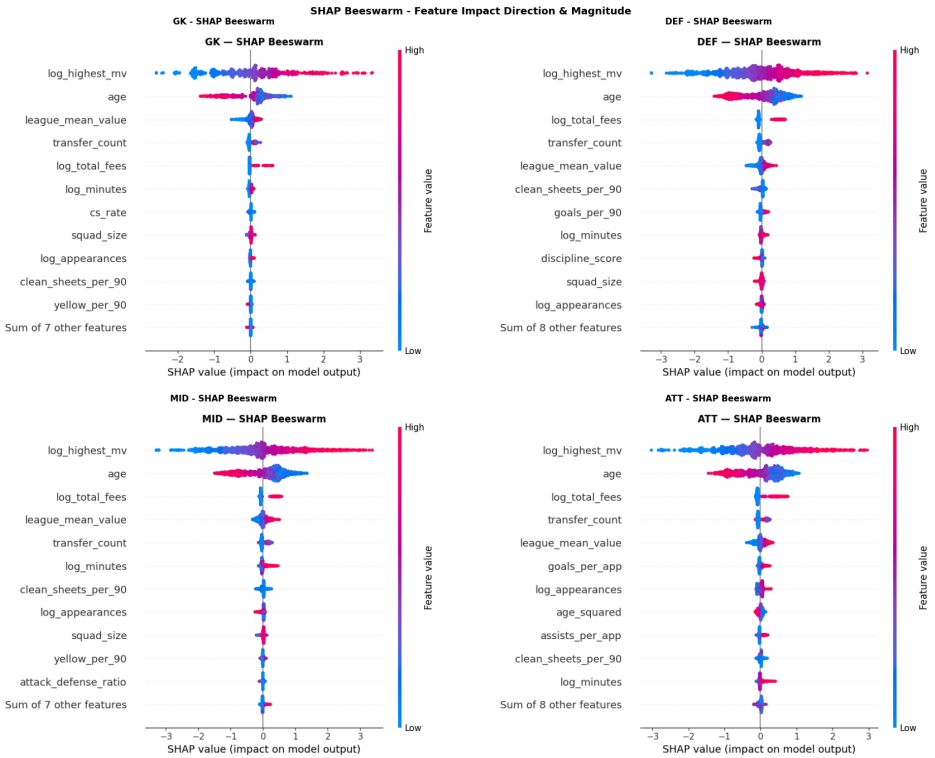
SHAP: What Drives Value?

Cross-position findings

- `league_mean_value` — top-3 in every position: *where* you play matters as much as *how*
- `log_highest_mv` — market anchors to career reputation
- `age²` — value peaks mid-20s, accelerates decline post-30

Position-specific signals

- **ATT**: `goals_per_90` becomes the strongest driver



Application: Dream Team

Scoring all 33,500+ players

Using serialised preprocessors — **no leakage**

Undervaluation ratio = predicted ÷ actual

Minimum 10 appearances filter applied

4-3-3 squad — cost vs. predicted value

Metric	Value
Total actual cost	**€0.50M**

Conclusions

What we achieved

- R^2 **0.79–0.83** across all positions
- XGBoost outperforms RF, MLP, and linear baselines
- SHAP provides actionable interpretation
- Full reproducible pipeline (10 notebooks)

Key insights

- League prestige = strongest cross-position signal
- Career peak value anchors market estimates
- Position-specific features add meaningful lift
- High-value players systematically underestimated (compression effect)

Future work

- Add injury history & international caps
- Temporal modelling (LSTM / Transformer) for career trajectory
- Multimodal: combine stats with video embeddings
- Real-time API for transfer window scouting

Thank You

Repository: `football-player-value-predictor`

Pipeline: 10 notebooks · 7 raw CSVs → dream team

Best model: XGBoost · R^2 up to **0.829** (MID)

Questions?