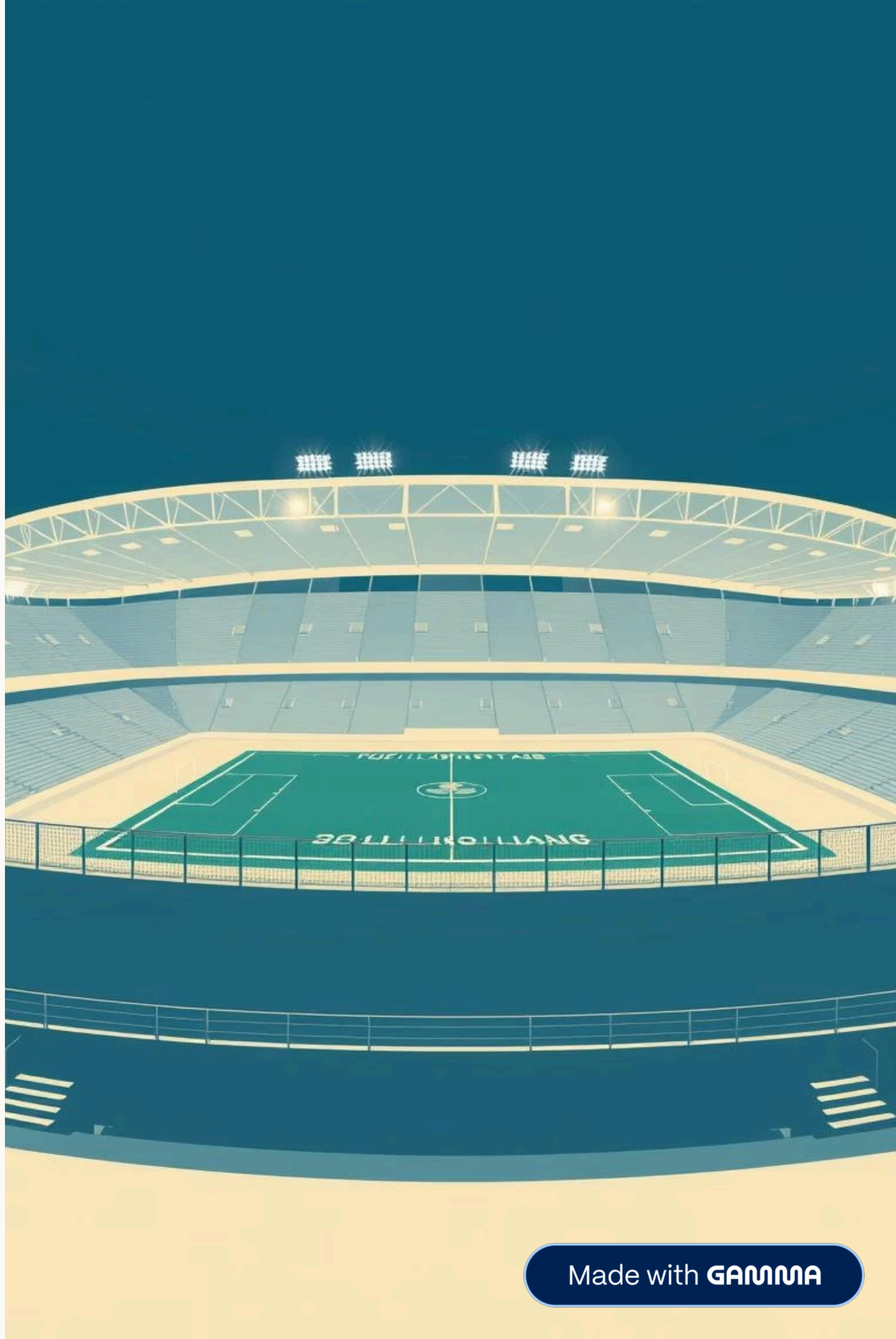
Predicting Future Football Talent Using Pre-18 Data

A Machine Learning Approach to Early Talent Identification in Professional Association Football.

Course: Data Mining



Problem Statement

The Complexity of Youth Scouting

- Identifying elite potential at the U18 level is fraught with high uncertainty due to non-linear physical and psychological maturation.
- Traditional scouting remains heavily reliant on subjective observation, leading to cognitive biases and "the Relative Age Effect".
- Performance data at youth levels is often sparse or fragmented across different regional leagues.

Actuality and Relevance

The strategic and financial imperative for early talent identification has never been greater in the modern game.

Financial Efficiency

Reducing the "transfer risk" by developing internal academy talent rather than purchasing established stars at a premium.

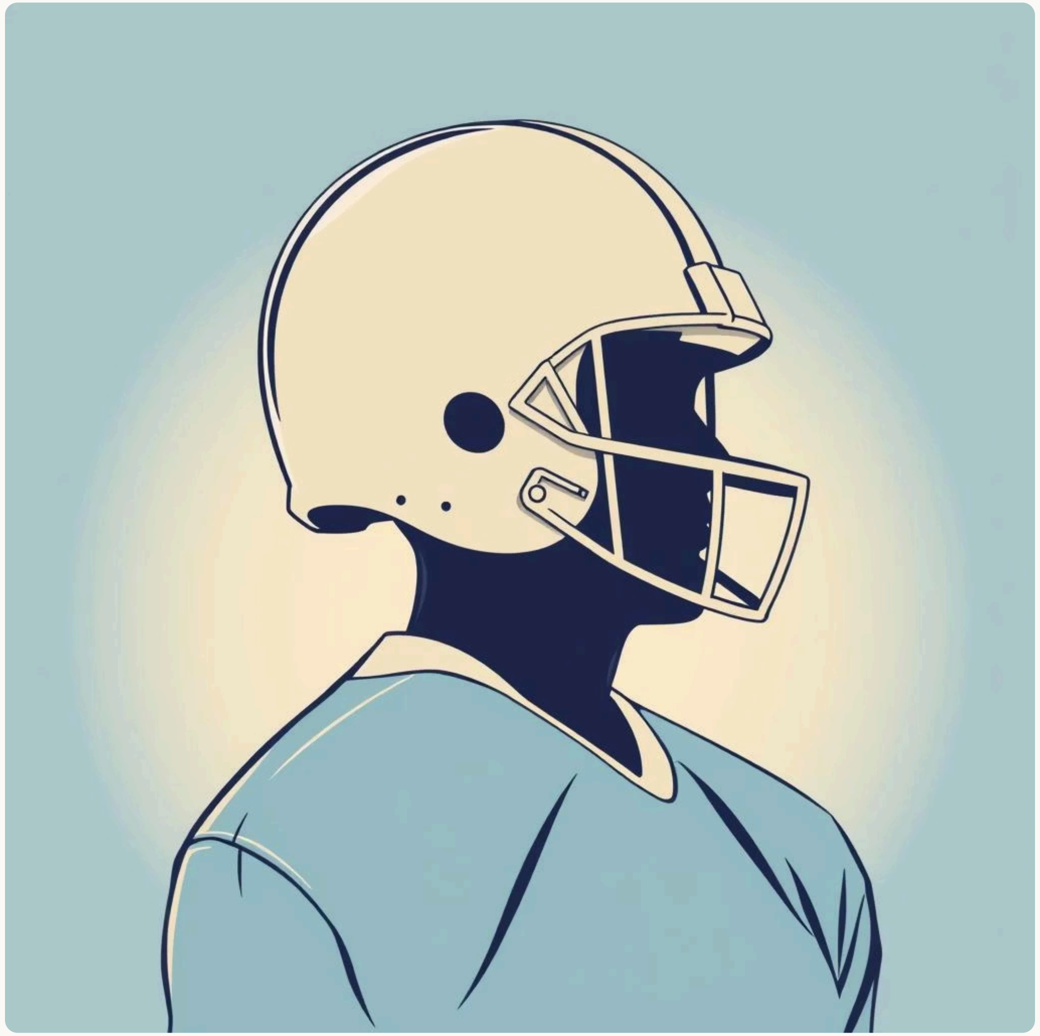
Strategic Planning

Enabling clubs to manage squad turnover years in advance by predicting which youth players will bridge the gap to the first team.

Resource Allocation

Directing elite coaching and medical resources toward players with the highest statistically backed probability of professional success.

Novelty and Originality



Methodological Contributions

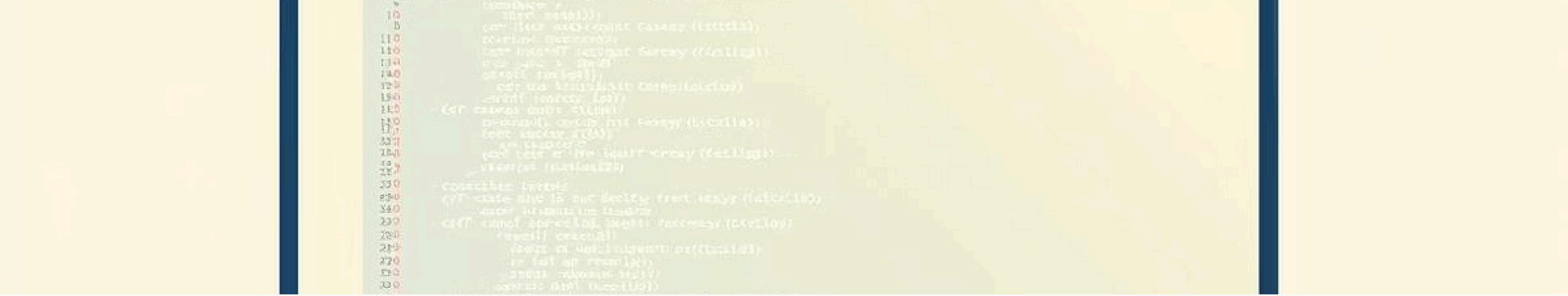
- **Pre-18 Constraint:** Strictly uses data generated before the player's 18th birthday to ensure true predictive value for academies.
- **Leakage-Safe Design:** Implements rigorous temporal splitting to prevent future performance data from influencing training sets.
- **Binary Success Definition:** Defines "Talent" based on objective market value and appearances in Top-5 European leagues by age 23.

Literature Overview

Current state-of-the-art systems primarily focus on similarity-based recommendations rather than long-term prediction.

Feature	Standard PSR Systems	Proposed Project
Focus	Current performance similarity	Future potential prediction
Data Scope	All ages / current season	Strictly Pre-18 developmental data
Goal	Find "the next Messi"	Identify "high-probability" pros

While existing research excels at identifying current top performers, it fails to account for the developmental trajectory essential for youth scouting.



Dataset Overview: Transfermarkt (Kaggle)

Scale

Global coverage of over 100,000+ players across multiple domestic and international tiers.

Granularity

Detailed tables including appearances, market value history, injuries, and competition rankings.

Temporal Depth

Historical data spanning over a decade, allowing for long-term longitudinal tracking of youth careers.

Preprocessing & Feature Engineering

01

Data Cleaning & Filtering

Removed players with incomplete historical records and filtered for those who had reached age 23 to establish ground truth.

02

Feature Aggregation

Aggregated youth stats (U15-U18) into a single feature vector representing the early career profile.

03

Core Engineered Features

- **Volume:** Total minutes played in senior vs. youth competitions.
- **Efficiency:** Goals and assists per 90 adjusted for league difficulty.
- **Consistency:** Number of distinct seasons with >500 minutes played.

Methods & Techniques

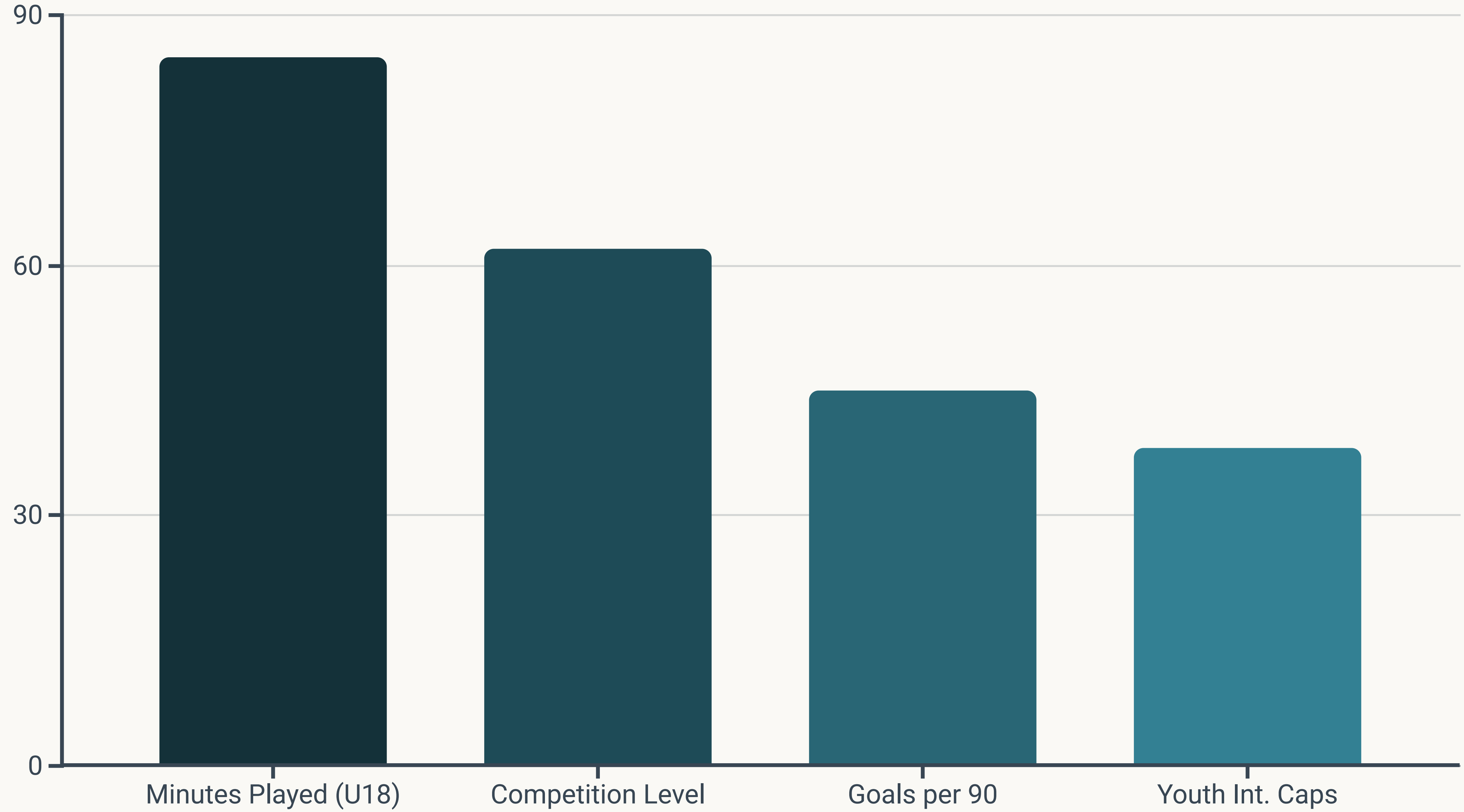
Model Selection

Supervised classification using gradient-boosted decision trees (GBDT) for their ability to handle non-linear relationships and missing values.

- **XGBoost:** For robust handling of sparse features.
- **LightGBM:** Utilised for efficiency and GOSS (Gradient-based One-Side Sampling).

Results & Key Metrics

LightGBM outperformed other models, achieving an AUC-ROC of 0.84 on the hold-out test set.



Finding: Early exposure to senior football (Minutes Played) is a significantly stronger predictor of elite success than raw goalscoring statistics at the youth level.

Conclusions & Future Work

Developmental Exposure

The quantity of playing time in competitive environments is the primary indicator of future professional viability.

Model Efficacy

Machine learning provides a reliable objective baseline that significantly reduces the noise inherent in youth scouting.

Proposed Next Steps

- **Spatial Tracking:** Incorporating GPS and event data to move beyond aggregated stats.
- **Position Specifics:** Developing tailored models for goalkeepers and defenders who mature later.
- **Explainability:** Using SHAP values to explain individual player predictions to coaches.