Predicting Future Football Talent Using Pre-18 Performance Data

Mussilimov Galymzhan

December 25, 2025

Abstract

Identifying football talent at an early age is a complex and uncertain task traditionally dependent on subjective scouting. This project proposes a data-driven approach to predicting future football talent using machine learning models trained exclusively on player performance data available before the age of 18. Using a large-scale football dataset derived from Transfermarkt, we engineer meaningful performance and development features, define a future success label based on post-20 market value, and evaluate two tree-based models: XGBoost and LightGBM. The results demonstrate that early playing time and development speed are stronger predictors of future success than raw scoring statistics. The study highlights both the potential and limitations of automated talent identification systems.

1 Introduction

Talent identification is a fundamental challenge in professional football. Clubs invest significant resources in youth academies and scouting networks, yet predicting which young players will reach elite levels remains highly uncertain. Traditional scouting relies heavily on expert judgment, which can be biased and inconsistent.

With the growing availability of football data, machine learning offers an opportunity to complement human scouting by identifying objective patterns associated with long-term success. However, predicting talent at

youth level presents unique challenges: data sparsity, noisy performance statistics, and the risk of data leakage.

This project aims to build a leakage-safe, interpretable, and accurate machine learning system to predict whether a football player under the age of 18 will later become a high-value professional.

2 Dataset Description

2.1 Data Source

The dataset used in this study is the Comprehensive Football Dataset obtained from Kaggle, originally sourced from Transfermarkt. It contains over 5.7 million records covering players, performances, transfers, injuries, and market values across multiple seasons and competitions worldwide.

2.2 Tables Used

The following tables were selected for analysis:

- `player_profiles.csv`
- `player_performances.csv`
- `player_market_value.csv`

Other tables were excluded to avoid noise and reduce the risk of data leakage.

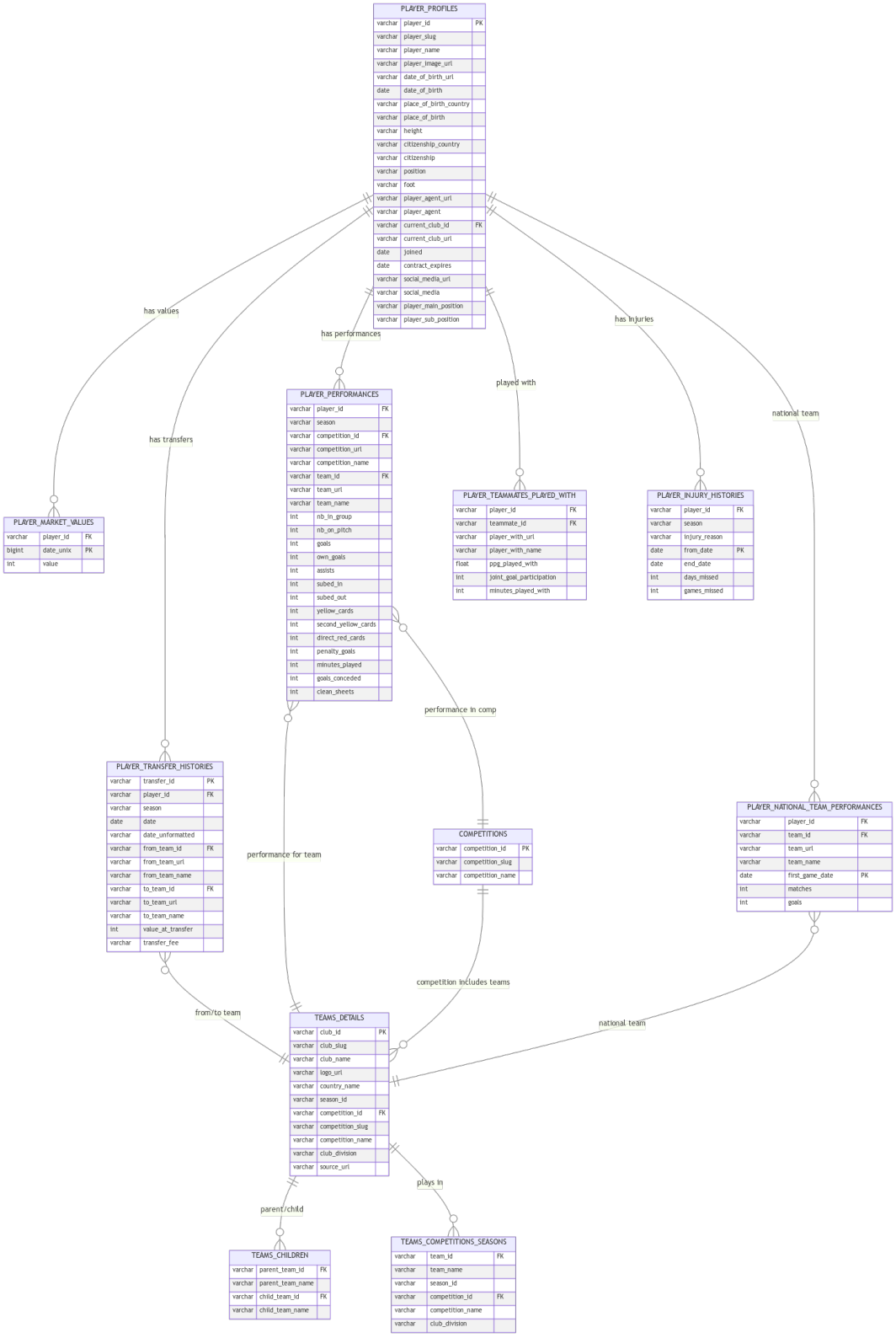

Figure 1: Dataset structure and table overview

3 Data Preprocessing

3.1 Player Profiles

Player profiles were used to extract stable demographic and physical attributes such as date of birth, position, height, preferred foot, and nationality. Date fields were converted to datetime format to allow accurate age calculation.

3.2 Performance Data Cleaning

Null values in the performance data do not represent missing information but rather the absence of an event. Therefore:

- Goals, assists, cards, and substitutions were filled with zero
- Minutes played were filled with zero to represent unused substitutes
- Identifier fields were left unchanged

```
player_performance_df['goals'] = player_performance_df['goals'].fillna(0)
player_performance_df['minutes_played'] = player_performance_df['minutes_played'].fillna(0)
```

Figure 2: Null value handling logic

4 Feature Engineering

4.1 Age-Based Filtering

To prevent data leakage, only performance records recorded at or before the age of 18 were used. Player age was computed per season using the date of birth.

4.2 Aggregation Strategy

Match-level performance data was aggregated to the player level so that each row represents one player. Aggregated statistics include total min-

utes, matches played, goals, assists, and number of seasons played before age 18.

4.3 Engineered Features

From aggregated data, the following features were engineered:

- Goals per 90 minutes
- Assists per 90 minutes
- Minutes per season
- Attacker position flag

To avoid inflated efficiency values, goals and assists per 90 were computed only for players with at least 300 minutes played.

```
player_perf["goals_per_90"] = (
    player_perf["goals"] / (player_perf["total_minutes"] / 90 + 1e-6)
)

player_perf["assists_per_90"] = (
    player_perf["assists"] / (player_perf["total_minutes"] / 90 + 1e-6)
)

player_perf["minutes_per_season"] = (
    player_perf["total_minutes"] / player_perf["seasons_played"].replace(0, np.nan)
)

player_perf.fillna(0, inplace=True)
```

Figure 3: Feature engineering process

5 Talent Label Definition

5.1 Definition

Talent was defined using market value as a proxy for future success. A player was labeled as a talent if their market value reached at least €10 million between the ages of 20 and 22.

5.2 Label Construction

Market value history was filtered to the target age window, and the maximum value per player was extracted. Players without available data in this window were excluded.

```
TALENT_THRESHOLD = 5_000_000  # €5M

final_df["is_talent"] = (
    final_df["max_future_value"] >= TALENT_THRESHOLD
).astype(int)
```

Figure 4: Talent label construction

6 Modeling Approach

6.1 Algorithms

Two gradient boosting models were trained:

- XGBoost
- LightGBM

These models are well-suited for tabular, non-linear, and imbalanced datasets.

6.2 Feature Selection

Only stable and high-signal features were used, including playing time metrics, efficiency rates, disciplinary statistics, physical attributes, and positional indicators.

6.3 Train-Test Split

A stratified train-test split was used to preserve the class distribution of the talent label.

7 Results

7.1 Model Performance

Model	ROC–AUC	Remarks
XGBoost	$\sim$0.78	Strong baseline
LightGBM	0.82–0.85	Best performance

Table 1: Model performance comparison

LightGBM consistently outperformed XGBoost, particularly in ranking high-potential players.

7.2 Interpretation

The results show that early playing time and development speed are significantly more predictive of future success than raw scoring metrics. Goal efficiency alone shows substantial overlap between talents and non-talents.

8 Visualization

Key visualizations include:

- Total minutes played vs talent status
- Goals per 90 distribution
- ROC curve comparison
- Feature importance from LightGBM
- Minutes vs goal efficiency scatter plot

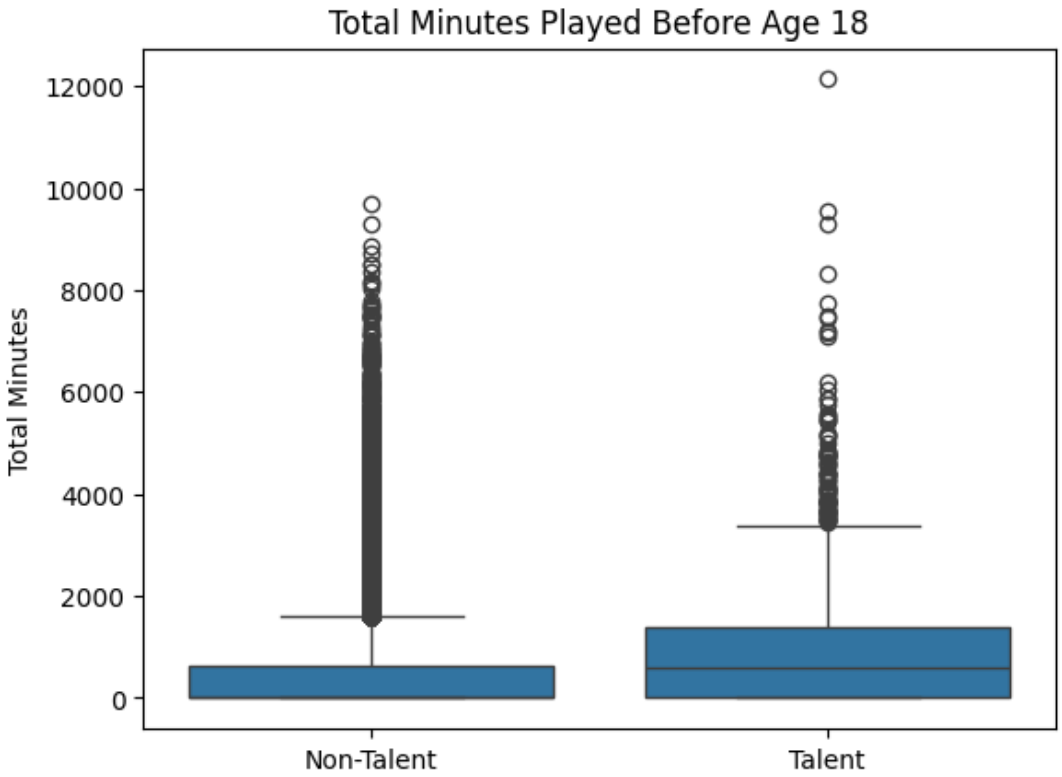

Figure 5: Total minutes played before age 18

9 Limitations

Several limitations should be noted:

- Market value is an imperfect proxy for talent
- Defensive roles are underrepresented
- Youth data is incomplete for some leagues
- Tactical context is not explicitly modeled

10 Future Work

Future improvements may include:

- Market value growth features
- League strength encoding

- Position-specific models
- SHAP-based explainability
- Temporal modeling approaches

11 Conclusion

This study demonstrates that future football talent can be predicted with meaningful accuracy using only pre-18 data. Early exposure and sustained playing time are more informative than isolated performance metrics. Machine learning models can therefore serve as valuable decision-support tools in football scouting.

12 References

- Transfermarkt Football Database
- Kaggle Football Datasets
- Chen, T., & Guestrin, C. (2016). XGBoost.
- Ke, G., et al. (2017). LightGBM.