Kazakh-British Technical University

School of Information Technology and Engineering

Predicting Future Football Talent Using Pre-18 Performance Data

Course: Data Mining / Machine Learning

Written Report

Authors:
Mussilimov Galymzhan
Tolegen Nursultan
Kasym Talgat

Instructor:
Adilet Yerkin

Almaty, Kazakhstan
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, meaningful performance and development features are engineered, a future success label is defined based on post-20 market value, and two tree-based models, XGBoost and LightGBM, are evaluated. The results demonstrate that early playing time and development speed are stronger predictors of future success than raw scoring statistics.

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 such as data sparsity, noisy performance statistics, and the risk of information 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 and originally sourced from Transfermarkt. It contains more than 5.7 million records covering player profiles, seasonal performances, transfers, injuries, and historical market values across multiple leagues and countries.

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 reduce noise and prevent data leakage.

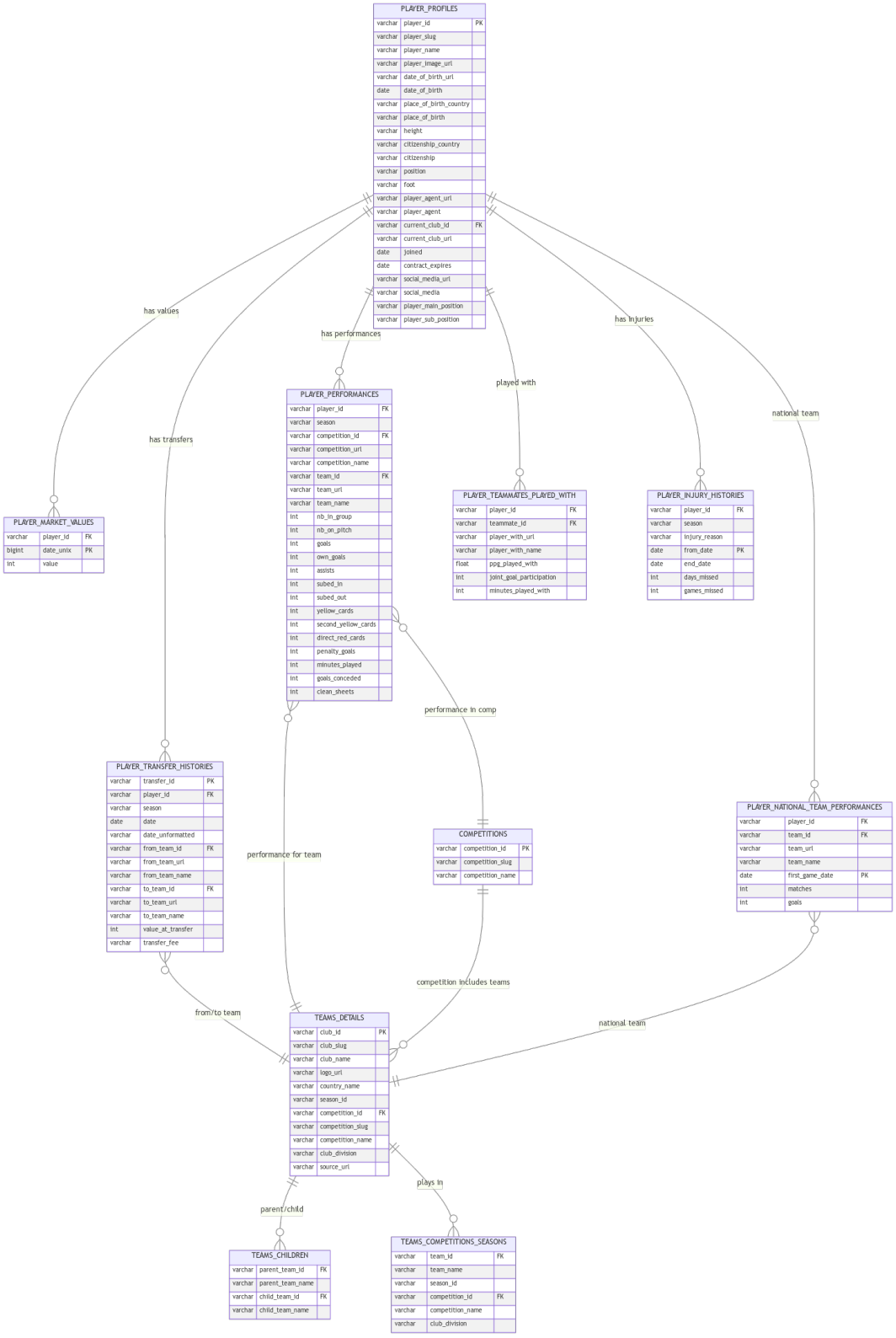

Figure 1: Dataset structure and table overview

Figure 1 illustrates the overall structure of the football dataset used in this study. Player profiles act as the central entity and are connected to

performance statistics and market value history through a unique player identifier. This structure enables a clear separation between early-career data used for feature extraction and future data used only for labeling, which is essential for preventing information leakage.

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 into datetime format to ensure accurate age computation.

3.2 Performance Data Cleaning

Null values in the performance data do not represent missing information but instead indicate the absence of an event. Therefore, football-specific cleaning rules were applied:

- Goals, assists, cards, and substitutions were replaced with zero
- Minutes played were set to zero for 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

Figure 2 demonstrates the logic used to handle null values in the performance table. Unlike traditional datasets, null values in football event data often indicate that an action did not occur. Replacing such values with zeros preserves the semantic meaning of the data and avoids distortions caused by statistical imputation.

4 Feature Engineering

4.1 Age-Based Filtering

To prevent information leakage, only performance records recorded at or before the age of 18 were retained. Player age was calculated 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 in the final dataset represents a single player. Aggregated statistics include total minutes played, matches played, goals, assists, and number of seasons played before age 18.

4.3 Engineered Features

The following features were engineered from aggregated data:

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

Efficiency metrics were calculated only for players with at least 300 minutes played to avoid inflation caused by small sample sizes.

```python
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

Figure 3 illustrates the feature engineering pipeline. Raw match-level statistics were transformed into stable player-level indicators that reflect development speed, efficiency, and early trust from coaches.

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 from training.

```python
TALENT_THRESHOLD = 5_000_000  # €5M

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

Figure 4: Talent label construction

Figure 4 shows the process used to construct the talent label. Future market value data was used exclusively for labeling, ensuring that no future information leaked into the feature set.

6 Modeling Approach

6.1 Algorithms

Two gradient boosting algorithms were used:

- XGBoost
- LightGBM

Both 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 indicators, disciplinary statistics, physical attributes, and positional flags.

6.3 Train-Test Split

A stratified train-test split was applied to preserve the 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 indicate that early playing time and development speed are significantly more predictive of future success than raw scoring statistics. Goal efficiency shows substantial overlap between talents and non-talents, highlighting its limited standalone value at youth level.

8 Visualization

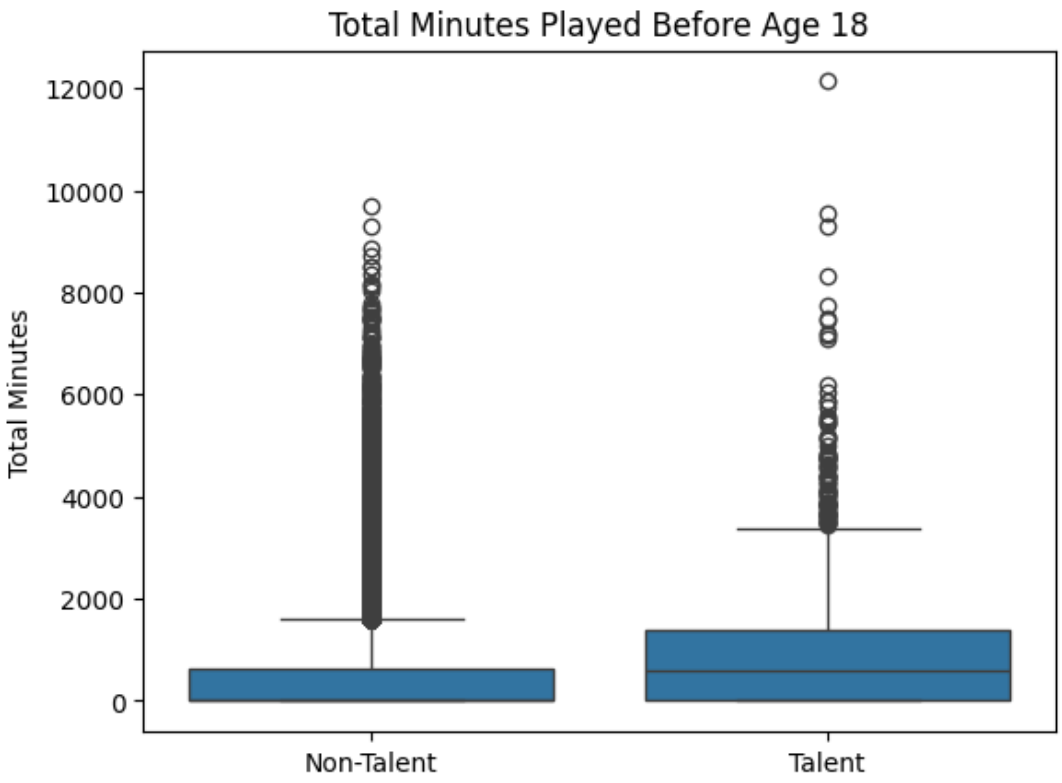

Figure 5: Total minutes played before age 18

Figure 5 demonstrates that players who later become talents accumulate significantly more playing time before age 18, reflecting early trust and consistent involvement.

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.