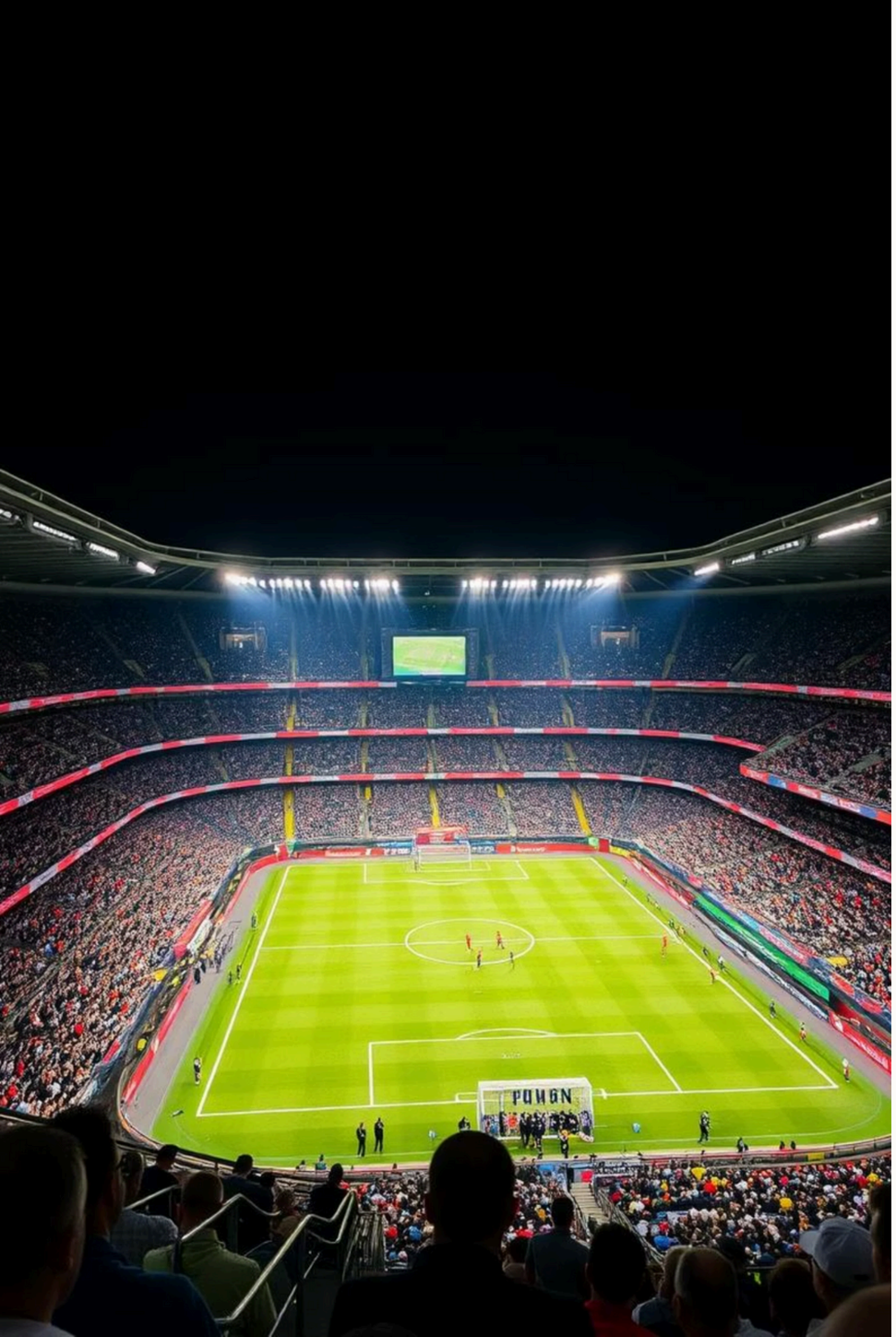


Football Players Market Value Prediction

Data Management Project - University of Milano Bicocca

Federico Cesare Cattò, Andrea Matteo Re, Chiara Pelizza

Project Introduction

Project Objective

To collect, integrate, and analyze football data to support data-driven decisions for team management.

Focus Area

Serie A club players — analyzing full performance across all competitions in the 2024/2025 season

Core Principle

Transition from traditional scouting to a sophisticated, analytical approach for competitive advantage.

Strategic Objectives

Data-Driven Player Evaluation

Analyze detailed performance metrics to objectively assess player contributions.

Market Value Estimation

Predict player market values using machine learning models based on statistical performance.

Underrated & Overrated Detection

Identify players whose market value significantly differs from model predictions, revealing hidden gems or overvalued profiles.

Squad Optimization Hypotheses

Develop and support hypotheses on team selection strategies and squad development.

Project Pipeline

Data Acquisition

Collect player statistics, market values, and contract information.

Data Integration & Enrichment

Merge diverse sources; enrich with team context and contractual data.

Data Preparation & Cleaning

Impute missing values, standardize variables, and remove inconsistencies.

Data Quality Assessment

Validate dataset completeness, accuracy, and consistency before modeling.

Data Storage

Store clean, structured dataset in SQL relational database for efficient querying and scalability.

Exploratory Data Analysis

Analyze trends and correlations for model guidance.

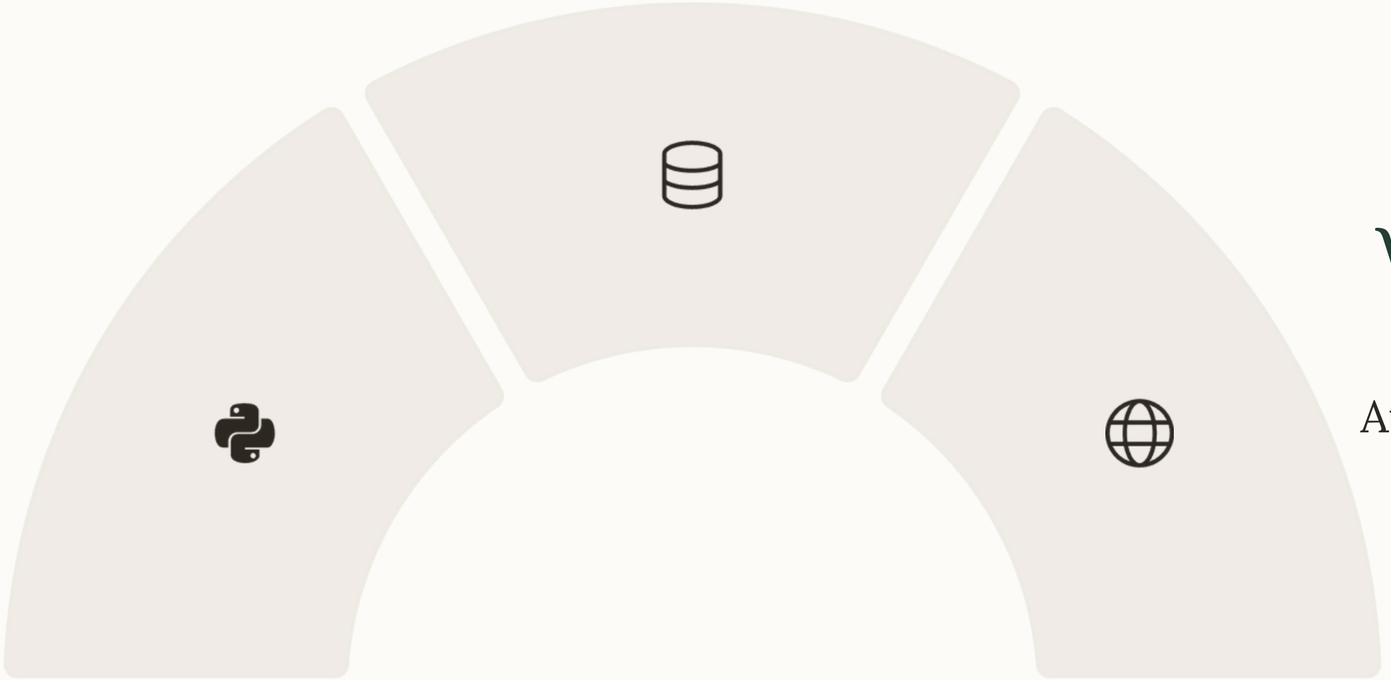
Predictive Modeling

Train neural network to estimate player market values and detect value gaps.

Tools and Technologies

SQL Database

Structured storage & efficient querying



Python

Key libraries: pandas, numpy, scikit-learn, matplotlib, seaborn

Web Scraping & APIs

Automated acquisition of online data

Data Acquisition Techniques

Web Scraping

Automated technique to extract data directly from websites.
A script reads the HTML content, identifies relevant information, and converts it into structured datasets.

API

Official interfaces that allow access to real-time, structured data (typically in JSON format) without interacting directly with web pages.

Our Data Acquisition Approach

Web Scraping

Data Sources:
✓ fbref.com[1] — Individual performance data (e.g., goals, assists, appearances)
✓ transfermarkt.it[1] — Market values

Automated extraction with BeautifulSoup & Selenium

API

✓ Football-Data.org to collect reliable player information and contract information (start/end dates)

Managed API rate limits with timed delays for stable data collection

[1] Links point to Atalanta; analogous sources were used for each Serie A club.

Data Integration & Enrichment

Key Strategies

- Player name standardization
- Outer joins to retain players
- Automated team processing
- Unified column names

Benefits

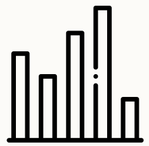
- Unified technical, market, contract data
- Supports advanced analyses
- Enables scouting & transfers

Challenges & Solutions

- Name inconsistencies → normalized
- Incomplete data → manual fixes
- API rate limits → timed delays

Data Sources & Initial Tables

Raw data collection from trusted football platforms



Fbref.com: Player Performance

Goals, assists, minutes played, and advanced metrics for individual player statistics.

Transfermarkt.it: Market Value

Player market values in euros.

Football-Data.org API: Contract Info

Team compositions and essential player contract dates.

Each source provided independent raw datasets requiring cleaning, harmonization, and integration.

Data Cleaning – Market Value & Contract Data

Standardizing financial and contractual information

Market Value Standardization

- Removed currency symbols and diverse text formats (e.g., "€3.5 mln", "€500 mila").
- Converted all collected market values to a unified numeric euro format for consistent analysis.
- Assigned null to unavailable or missing entries.

Contract Data Processing

- Extracted contract end year from text fields (e.g., "2026-06" → 2026).
- Computed remaining contract duration relative to 2025 for each player.

Data Merging for Complete Dataset

From fragmented sources to a unified Serie A players dataset

Merging Diverse Sources

- Merged datasets by player name with left join.
- Retained all performance records, added market values when available.
- Tracked data loss from mismatches

Concatenation of Teams

- Combined all cleaned team dataframes into one Serie A dataset.
- Ensured consistent columns across all clubs.

Data Imputation & Standardization

Imputation of Missing Values

Team: Filled using rule-based logic and manual validation.

Years of Contract Remaining: Estimated from public records and transfer info.

Position: Inferred from match data and player profiles.

Nation: Completed via external databases and official records.

Data Standardization

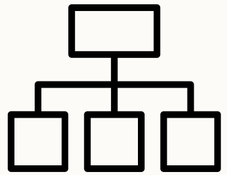
Age: Extracted minimum age from ranges (e.g., “25-096” → 25) and converted to numeric.

Nation: Extracted uppercase country codes (e.g., “en ENG” → “ENG”).

Position: Normalized multi-role entries by splitting, keeping primary role, and standardizing (e.g., “DF-MF” → “DF”).

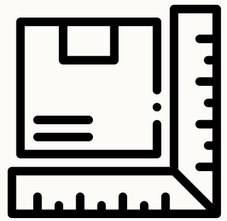
→ Dataset ready for analysis and modeling! ←

Dataset Description



Structure

Our dataset provides a comprehensive view of player performance, contracts, and market values for all Serie A players (2024-2025 season).



Dataset Size

209 variables: Comprehensive coverage including detailed performance metrics, contract specifics, and market valuation data.

Over 500 unique players: Analyzed from the Serie A 2024-2025 season, providing a broad and deep talent pool for analysis.

Dataset Description

Main Variable Categories

Individual Performance Statistics

Key quantitative indicators: goals, assists, passing accuracy, defensive contributions, minutes played.

Market Value (EUR)

Standardized player market value in euros, sourced from trusted platforms.

Remaining Contract Duration (years)

Years left on player contracts, computed from contract end dates.

Club Identifier

Team affiliation for each player to contextualize performance and value.

Data Quality – Quantitative Overview

545
Unique Players
Serie A 2024/25 Dataset

377
Manual Imputations
Ensuring data completeness

17
Outdated Recors
Succesfully removed

3
Implausible Ages
Corrected for accuracy

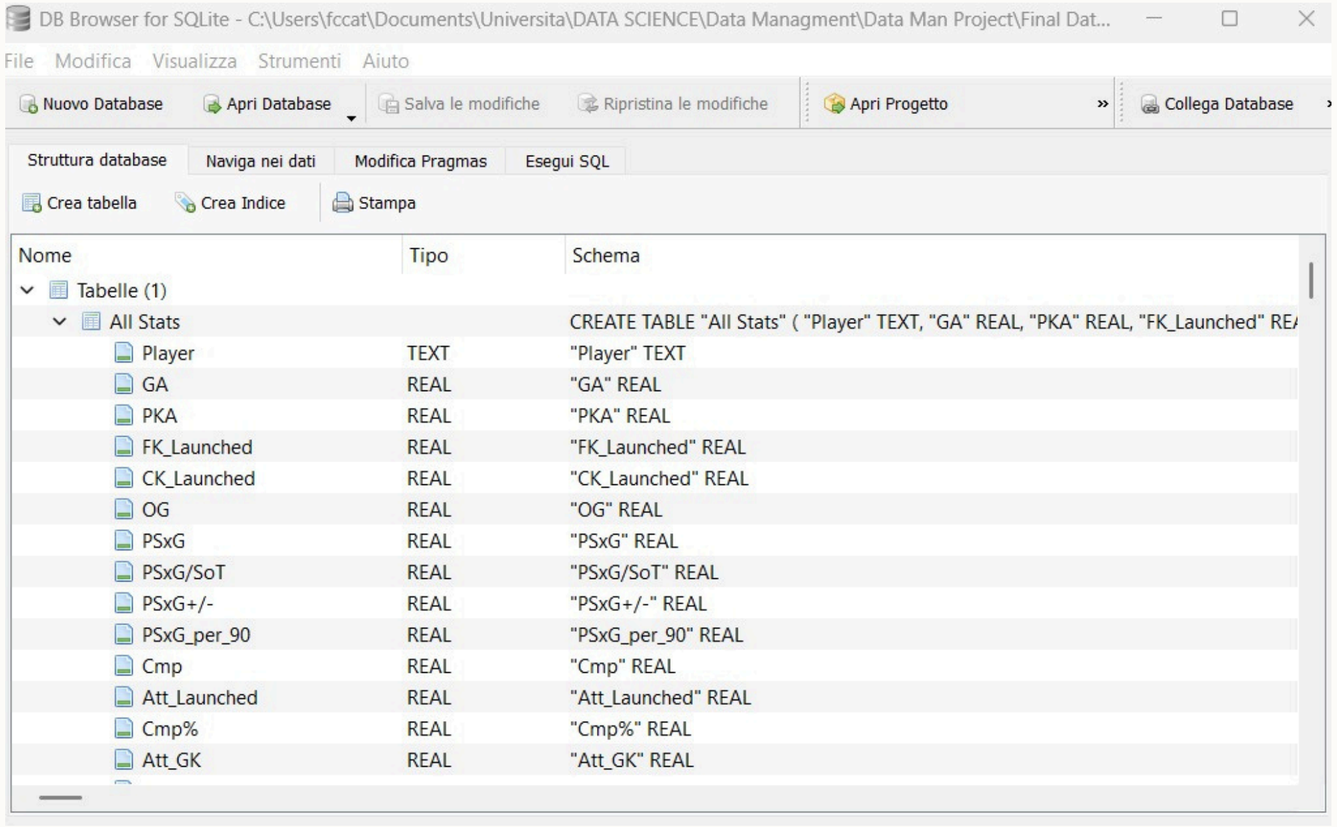
34
Duplicates Removed
Approximately 6.2% reduction

62.17%
Dataset Completeness
Overall data fill rate

100%
Consistency & Accuracy
Verified data integrity

Data Storage Solution

Data stored in a structured relational database using SQL



Why SQL for Data Storage?

Declarative Syntax

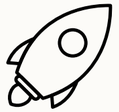
Simplifies complex queries, allowing focus on what data is needed, not how to retrieve it.

Data Integrity

Ensures consistent and reliable data management with robust ACID principles.

Standardized Language

Utilizes ANSI SQL for universal compatibility across diverse database platforms.



Scalability

Efficiently handles large datasets through advanced indexing and partitioning techniques.

Seamless Integration

Connects effortlessly with Python, R, and various Big Data ecosystem tools.

Optimal for Structured Data

Ideal for relational, analysis-ready datasets, offering strong schema enforcement.

Why a Single Unified Table?

Semantic Clarity

NULLs distinguish "Not Applicable" from real zeros.

ML Optimization

Ideal for tree-based models like CatBoost and XGBoost.

Simplified Workflow

Streamlines feature engineering and uniform data access.

Global Interpretability

NULL fields support a single, adaptable model suitable for deployment.

Data Integrity

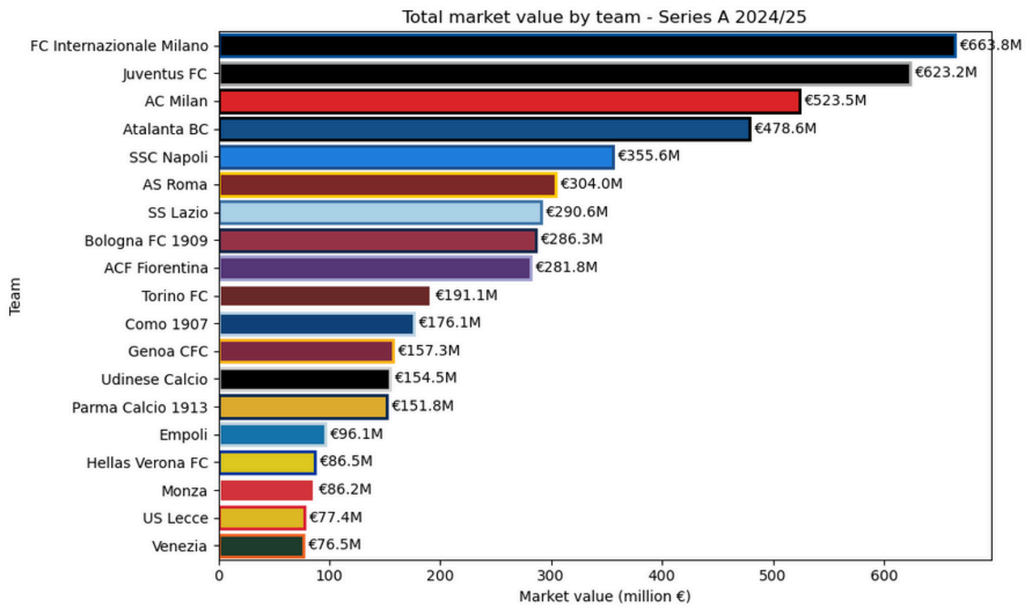
Prevents false patterns and data leakage.

ML Best Practices

Ensures reliable and robust predictive models.

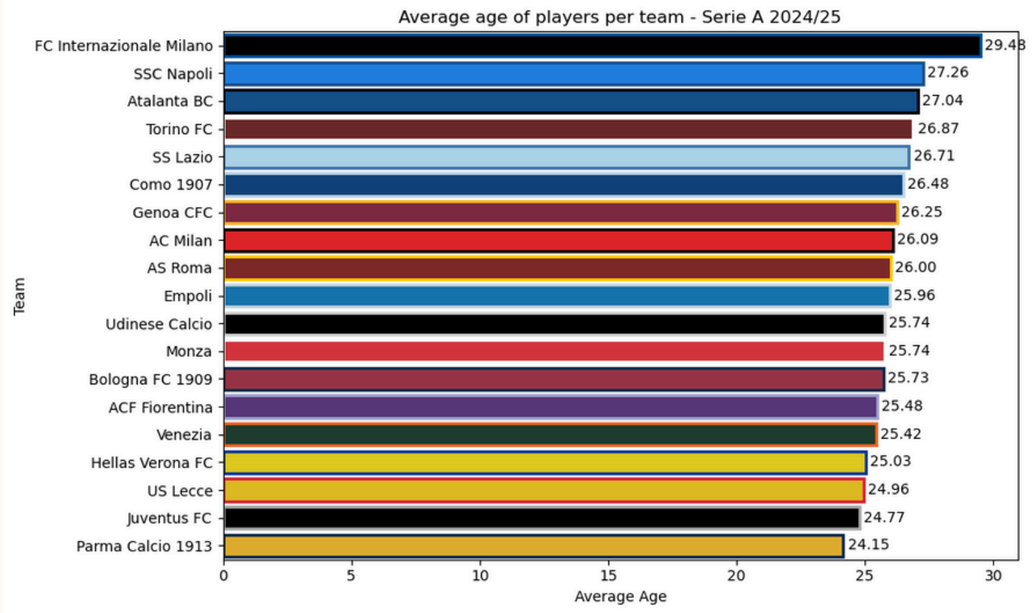
General Overview of Serie A 2024-25

Total Market Value by Team



- Wide gap in total market values across Serie A clubs
- Financial resources impact competitiveness, transfers, and team quality

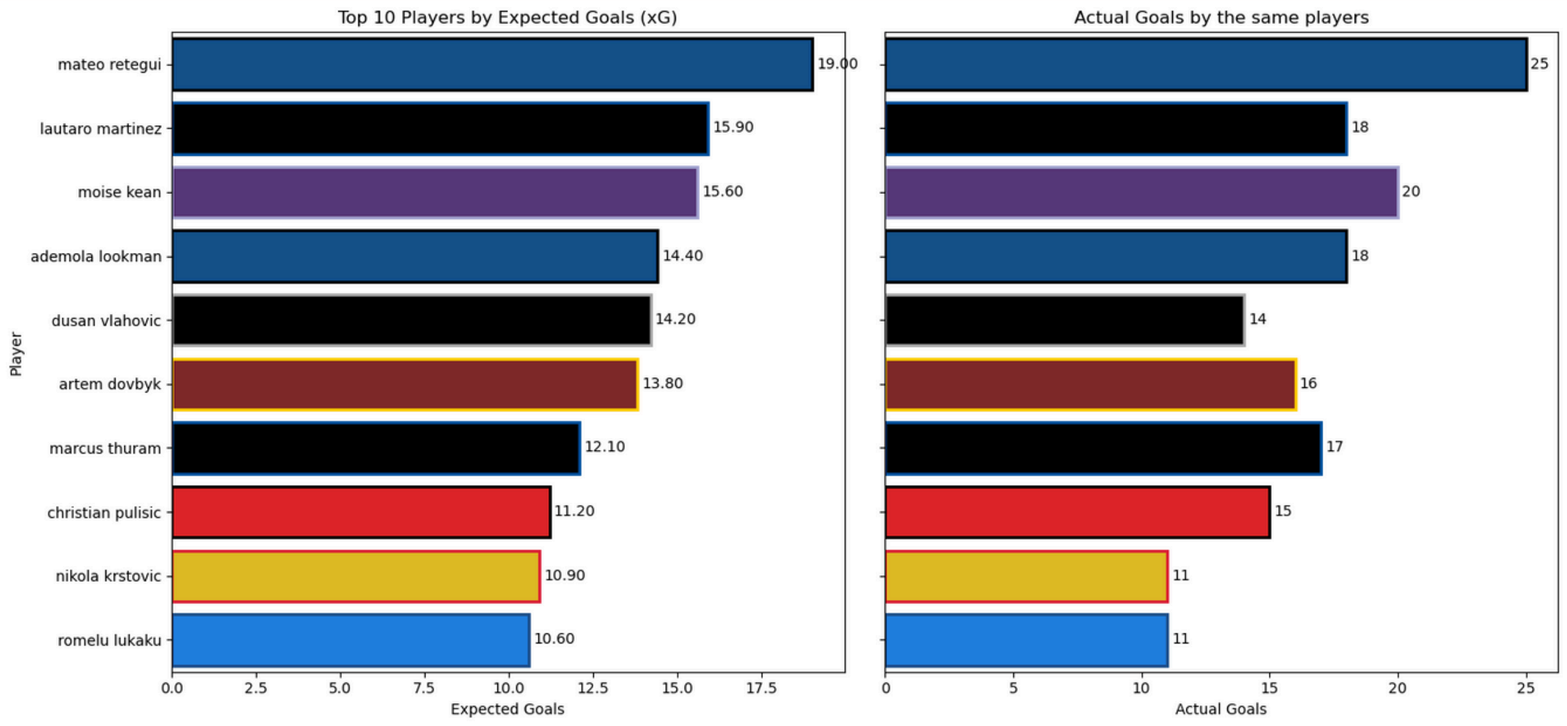
Average Squad Age by Team



- Team age reflects each club's strategic approach
- Younger teams bring dynamism and growth, while older squads focus on discipline and results.

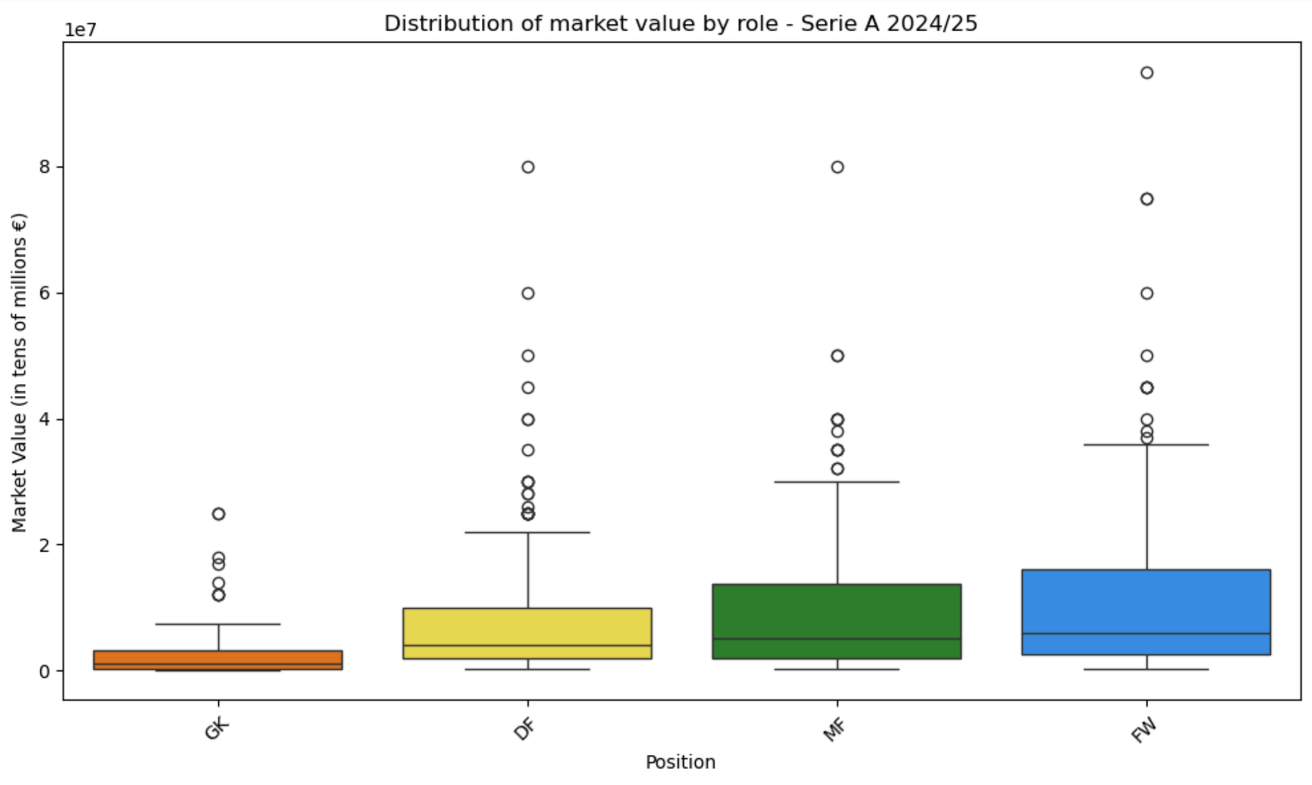
Player Scoring Efficiency

Expected Goals (xG) vs Actual Goals

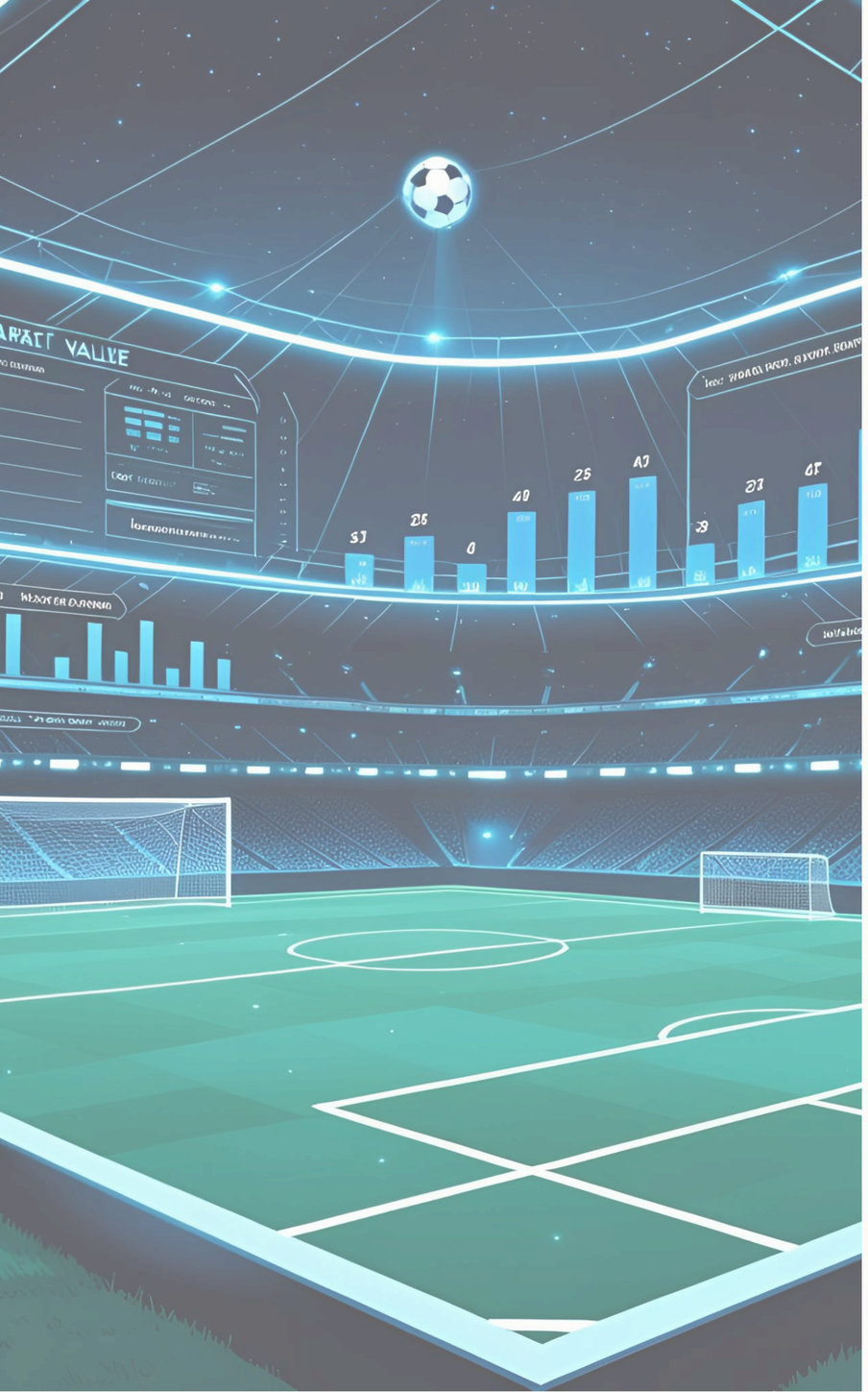


- Overperformers convert difficult chances efficiently
- Underperformers score below expected levels, indicating inefficiencies
- xG analysis reveals finishing quality, tactical roles, and conversion ability

Market Value Distribution by Role



- Goalkeepers show the most compact valuation range, with rare exceptions
- Defenders have moderate values, with some high-value outliers
- Midfielders show broad value distribution, reflecting hybrid tactical roles
- Forwards have the highest median market value and most variability — attacking talent commands a premium



Predicting Football Market Value

Project Objective

Predict Serie A players' market value to support data-driven scouting and smarter transfer decisions

---

Algorithm Choice - CatBoostRegressor

- Handles missing data natively
- Differentiates "0" from "not available"
- Robust for mixed, real-world football data
- Leave-One-Out Cross-Validation for evaluation
- Logarithmic target transformation to stabilize variance

Predicting Football Market Value
Model Performance

€ 4,196,636 MAE

Average difference between predicted and actual values.

Main Observations

Underestimates high-profile players' value, influenced by global fame or contract clauses.

Overestimates lesser-known players' value, identifying potential undervalued transfer targets.

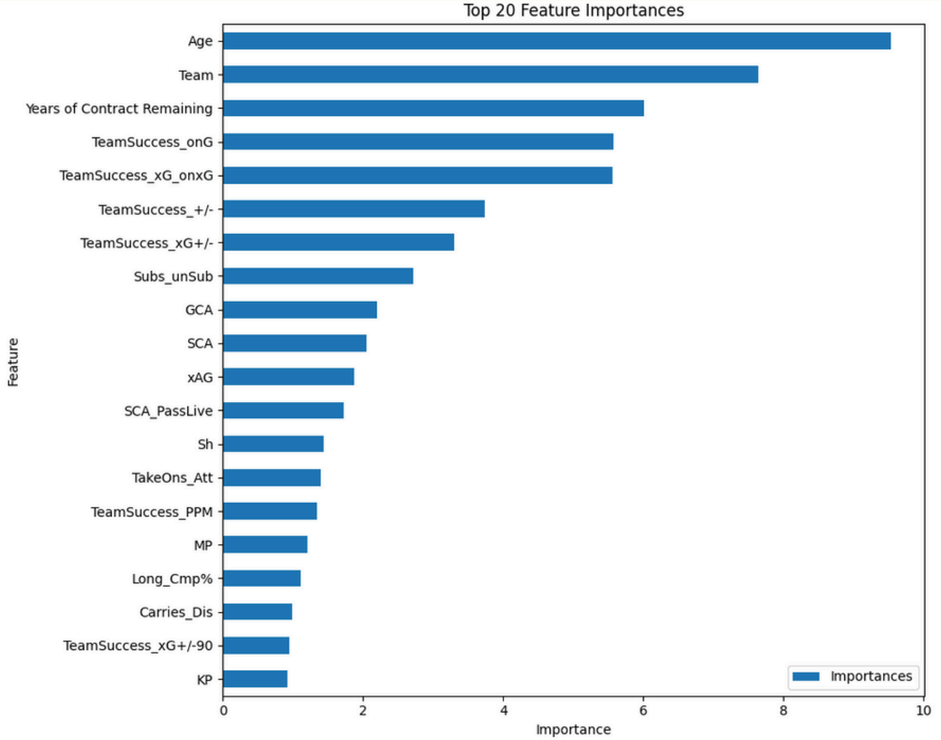
Overall, the model demonstrates solid performance, given the complex and volatile nature of the transfer market.

Predicting Football Market Value

Top 20 Features Impacting Market Value Prediction

Age, team success, and contract details drive value

Creative actions and playing time boost visibility



Market context and club prestige matter

Model reflects objective, performance-based drivers

Conclusions

- ✓ Quite accurate market value prediction for Serie A players
- ✓ Integrated and cleaned multi-source dataset
- ✓ High data quality ensured (completeness, consistency)
- ✓ Efficient data storage for ML
- ✓ Key insights on player and team market trends
- ✓ Solid foundation for data-driven transfer decisions

Future Developments

Dataset Enrichment

Incorporate injury history, fitness data, social media engagement, agent and contract info, and detailed match statistics.

Broader League Coverage

Expand beyond Serie A for more generalizable models and cross-league transfer analysis.

Advanced Modeling Techniques

Use ensemble models, deep learning, and probabilistic approaches for better prediction accuracy.

Automated Data Pipelines

Develop automated, monitored data collection with incremental updates for scalability and near real-time predictions.

Thank You
For The Attention