Football Players Market Value Prediction

Federico Cesare Cattò, Andrea Matteo Re, Chiara Pelizza
Università degli Studi di Milano Bicocca

Abstract

The project follows the key stages of a structured data management process.

Data acquisition involved collecting both structured and unstructured information through web scraping (using BeautifulSoup and Selenium) and APIs. We extracted detailed statistics from sources such as Transfermarkt and FBref.com, and complemented them with contractual details from API.football-data.org. These heterogeneous data sources were integrated and enriched into a unified dataset using unique player identifiers and outer joins, combining performance metrics, economic values, and contract information.

Data preparation included standardizing market values, processing contractual data, imputing missing values, recoding variables and normalizing player positions.
A detailed data quality assessment was conducted to enhance completeness, consistency and accuracy.
For persistent storage, a relational SQL database was employed, chosen for its ability to maintain data integrity, structured schema and compatibility with machine learning workflows.

Exploratory Data Analysis (EDA) was performed to identify trends, relationships and anomalies for predictive modeling. The core of the project was the development of a database for a CatBoostRegressor model to estimate players' market value based on their performance statistics. The model was trained and validated using Leave-One-Out Cross-Validation (LOOCV), achieving a mean absolute error (MAE) of approximately € 4.2 million on held-out players.

Results revealed that the model tends to underestimate high-profile players while overestimating lesser-known ones, indicating that intangible factors such as media attention and commercial appeal—absent from the included features—affect extreme market valuations. Feature importance analysis highlighted age, team affiliation, remaining contract length and team success metrics as the main drivers of predicted market value.

Overall, the project delivers a data-driven tool that can assist football clubs in identifying pricing inefficiencies and supporting more informed transfer strategies, providing an objective perspective on player valuation in professional football.

Contents

1 Introduction

In the modern era of sports analytics, the application of **data science** and **machine learning** techniques has significantly reshaped how players are analyzed, valued and scouted. The domain of footballand specifically **Serie A**—Italy's premier football league—offers a rich context for developing data-driven methodologies aimed at understanding and evaluating player performance.

Serie A features 20 competing teams each season. The clubs that took part in the 2024–2025 season are as follows: Atalanta[1] [2], Bologna[3] [4], Cagliari[5] [6], Como[7] [8], Empoli[9] [10], Fiorentina[11] [12], Genoa[13] [14], Hellas Verona[15] [16], Inter[17] [18], Juventus[19] [20], Lazio[21] [22], Lecce[23] [24], Milan[25] [26], Monza[27] [28], Napoli[29] [30], Parma[31] [32], Roma[34] [35], Torino[36] [37], Udinese[38] [39], Venezia[40] [41]

This project adopts a complete **data management** pipeline with the goal of predicting the **market value** of players based on performance statistics.[1]

The first phase of the project is focused on **data acquisition**, which involves gathering structured and unstructured data from various online sources. We employ both **web scraping techniques** and **APIs** to collect a wide range of statistics for Serie A players. These include traditional performance metrics (such as goals, assists and minutes played), as well as more advanced indicators related to passing, defending and physical contributions.

Once collected, this data is stored in a **persistent database**, ensuring long-term accessibility and consistency. During the **data storage** phase, the focus is on organizing the data into structured formats that facilitate efficient querying and analysis.

The next step involves **data integration and enrichment**, where multiple sources are merged to construct a comprehensive dataset. At this stage, we aim to enhance the original data by combining it with contextual information such as team performance, league position. This helps us gain a deeper and more nuanced understanding of each player's value and performance context.

Before progressing to modeling tasks, we dedicate a phase to assessing **data quality**. This involves evaluating and improving several key quality dimensions, including **completeness**, **accuracy**, **consistency** and **timeliness**. Ensuring high-quality data is critical to the integrity of subsequent analyses and predictions.

With a clean and enriched dataset, we conduct **exploratory data analysis (EDA)** to identify initial trends, relationships and potential anomalies in the data. This step involves querying the dataset to uncover valuable insights and guide feature selection for our predictive modeling. The EDA provides a better understanding of how various performance metrics correlate with players' market values and prepares the groundwork for model design.

The project lies in the implementation of a **CatBoostRegressor** aimed at estimating players' **market value** based on their statistical performance. We adopt a **supervised learning** approach, where the input features are drawn from the performance metrics collectedand the output variable is the **actual market value**, sourced from trusted platforms such as Transfermarkt. The model learns from this data to approximate the relationship between on-field

performance and financial valuation.

Following the prediction phase, we analyze the **differences between predicted and actual values**, which allows us to detect players who are statistically **overrated** or **underrated**. This comparison not only highlights market inefficiencies but also provides insight into potential biases in the current valuation system used in the football transfer market.

To support and communicate our findings, we introduce a set of **data visualizations** in the final part of the analysis, such as bar charts.

From a methodological standpoint, the structure of our project follows the key stages of a data management pipeline. These can be summarized as follows:

- **Data acquisition**: collection of relevant information through *web scraping* techniques and publicly available *APIs*;
- **Data integration and enrichment**: merging and enhancing datasets to improve completeness and insight;
- **Data preparation and cleaning**: transformation and sanitization of raw data to ensure consistency, usability and compatibility with analytical methods;
- **Data quality**: evaluation and refinement of data accuracy, consistency and completeness;
- **Data storage**: persistent storage of the retrieved data in a *structured database*;
- **Exploratory data analysis (EDA)**: querying and visualization to identify trends and guide model design;
- **Predictive analytics**: estimation of player value using neural networks, with the potential to propose an *optimal squad* for a future tournament based on statistical performance.

In the following chapters, each of these steps will be discussed in greater detail.
By leveraging structured data and modern machine learning techniques, our aim is to contribute a data-driven perspective on player valuation within professional football.

2 Data Acquisition

Data acquisition is a critical component of the *data science workflow*, serving as the **foundational step** upon which all subsequent analysis, modeling and interpretation depend. It refers to the process of **identifying**, **collecting** and **storing** data from various sources, ensuring it is **accurate**, **reliable** and **suitably structured** for analysis.

In the context of **sports analytics**and more specifically regarding the **Italian Serie A 2024-2025** season, robust data acquisition enables informed evaluations of *player performance*, *market value of players* and *contract informations*.

High-quality acquisition influences the **validity**, **depth** and **reproducibility** of analysis, making it essential to choose the right tools and techniques.

2.1 Methods of Data Acquisition

Several techniques are commonly employed:

- Web Scraping
- APIs
- Manual Collection

2.1.1 Web Scraping

Web scraping involves programmatically extracting data from websites. Two popular Python libraries are:

- **BeautifulSoup**: Ideal for *static HTML*. Lightweight and efficient for parsing.
- **Selenium**: Used for *dynamic websites* rendered with JavaScript. Simulates browser interaction.

2.1.2 APIs (Application Programming Interfaces)

APIs offer structured, reliable access to databases and typically return data in *JSON format*. They are efficient and well-documented.

2.1.3 Manual Collection

While not scalable, **manual collection** may be necessary when data is only available in *PDFs or tables*, not accessible via APIs or scraping.

2.2 Our Data Acquisition Process for Serie A Player Statistics (2024-2025)

Due to the lack of a unified dataset, we adopted a **hybrid approach** combining web scraping and API usage.

2.3 Sources Used

- **Transfermarkt** and **FBref.com** – for detailed statistics (scraped)
- **API.football-data.org** – for contract data (API)

2.4 Tools and Libraries

- **Python** – scripting
- **BeautifulSoup** – parsing static HTML
- **Selenium** – rendering JavaScript
- **Pandas** – dataframes and CSV output
- **Requests** – API and HTML retrieval

2.5 Scraping with FBref (BeautifulSoup + Selenium)

Selenium was used to *load dynamic content*, while the HTML source was passed to BeautifulSoup.
We extracted tables containing:

- **Standard Stats**
- **Goalkeeping** and **Advanced Goalkeeping**
- **Shooting**, **Passing**, **Defensive Actions**, **Possession**
- **Goal and Shot Creation**, **Playing Time**, **Miscellaneous Stats**

Tables were parsed using `pandas.read_html()` and saved as CSV files.

2.6 API Integration (football-data.org)

We accessed structured player data using RESTful API calls with token-based headers, retrieving:

- **Team metadata** and standings
- **Squad lists**
- **Player profiles**: nationality, birthdate, position, contract duration

2.7 Challenges Encountered

- **Rate limits** on APIs – solved with timed delays
- HTML **tables within comments** – required custom parsing
- **Name mismatches** – resolved via normalization

Our **multi-source**, **multi-method** strategy enabled the creation of a **rich, scalable** dataset for Serie A player analysis. The combination of scraping and APIs ensured **comprehensive**

coverage, while modular scripting supports easy updates and reusability for future seasons or similar leagues.

3 Data Integration and Data Enrichment

In the realm of data science and analytics, **data integration** and **data enrichment** are two fundamental yet distinct processes that significantly enhance the quality, usability and depth of analytical datasets. Even though they are often interrelated and may occur in tandem within data pipelines, it is important to recognize their conceptual differences and the unique value each contributes to the overall data lifecycle.

Data Integration

Data integration refers to the process of *combining data* from multiple heterogeneous sources into a unified, coherent dataset. This process typically involves resolving issues of **data format**, **schema alignment** and **semantic consistency**, enabling the analyst to treat multiple datasets as a single, logically unified source. Data integration is particularly vital in environments where information is distributed across disparate systems—such as databases, web platforms, APIs and flat files— and needs to be consolidated to support a unified analytical model.

The key benefits of data integration include:

- **Improved completeness** of data by merging diverse attributes and variables.
- **Enhanced data consistency** through normalization and schema matching.
- **Streamlined querying and analysis**, as data is available in a single integrated form.

Data Enrichment

Data enrichment refers to the process of enhancing existing datasets by integrating additional relevant information from external sources or through further processing. This technique plays a crucial role in modern data analysis and research because raw or initial datasets often lack the necessary context or variables to fully capture the complexity of the phenomena being studied. By enriching data, analysts and researchers can obtain a more complete, accurate and meaningful representation of the subject matter.

The primary purpose of data enrichment is to improve the quality and depth of the data, thereby enabling more robust analyses and reliable conclusions. For example, adding temporal, geographic or contractual information can reveal patterns and dependencies that would otherwise remain hidden. Enriched datasets often facilitate better predictive modeling, hypothesis testing and strategic decision-making.

Among the key strengths of data enrichment are:

- **Improved analytical accuracy**: incorporating additional relevant variables increases the explanatory power of statistical and machine learning models, leading to more precise predictions and insights.
- **Enhanced decision-making**: more comprehensive datasets provide decision-makers with richer information, allowing them to make better-informed and more confident choices.

- **Uncovering hidden relationships** by combining different data sources or adding derived features, enrichment can reveal correlations or causal links that remain obscured in the original data.
- **Greater dataset versatility**: enriched data becomes applicable to a broader range of research questions and use cases, expanding its usefulness beyond the initial scope.
- **Support for dynamic and longitudinal analyses**: adding time-related variables or contract details, for example, enables the study of changes and trends over time rather than static snapshots.

3.1 Data Integration and Enrichment in the Construction of a Serie A Players Dataset

The process implemented to build a comprehensive dataset for Serie A players across all teams for the 2024-2025 season exemplifies a well-structured pipeline of **data integration**. In this case, the integration consists in combining multiple heterogeneous data sources—each contributing distinct but complementary information—into a single, unified dataset.

3.1.1 Key Integration Strategies

- **Uniform Player Identifier**: integration is based primarily on the player name, which has been normalized using the `unidecode` package and string transformations to handle character encoding issues, different spellings and whitespace discrepancies. This standardization is critical for successful merges across dataframes.
- **Outer Joins and Missing Data Management**: outer joins are employed during merging to ensure that no potentially useful player information is lost due to missing entries in one dataset. After merging, the proportion of unmatched entries (e.g. players lacking market value data) is quantified to assess the integration completeness.
- **Team-Based Aggregation and Loop Automation**: the same procedure is repeated for all Serie A teams. Scripts are modular and structured to allow looping through each team, pulling its data and writing standardized CSV files, which are then concatenated into a single dataframe.
- **Semantic Consistency**: column names across datasets are renamed to follow a coherent naming convention. For instance, columns like `"Valore di Mercato"` from Transfermarkt are renamed to `"Market Value"` and contract information fields are aligned across all teams.
- **Final Consolidation**: after enriching the Atalanta dataset with contract and market data, the same approach is replicated for all other Serie A teams. The final integration step involves concatenating all individual team dataframes vertically into one master dataset: `Serie_A`.

3.1.2 Benefits of the Integration

This process demonstrates the power of data integration to:

1. **Unify diverse information**: by combining technical performance, economic value and contractual data, a multi-dimensional perspective of each player is achieved.

2. **Enable richer analyses**: this enriched dataset allows for complex queries, such as the relationship between contract duration and performance metrics or correlations between market value and in-game statistics.
3. **Support decision-making**: for clubs, analysts and stakeholders, the integrated data provides a holistic view useful for scouting, transfer negotiations and salary budgeting.

3.1.3 Challenges Encountered

- **Name Disambiguation**: one of the most significant challenges in data integration is handling inconsistent naming conventions. Differences due to accents, middle names or abbreviations can lead to incorrect or missed merges.
- **Data Availability**: some teams (e.g. Monza, Venezia, Empoli) lacked complete data from certain sources, requiring either manual intervention or exclusion from specific enrichments.
- **Rate Limits and API Constraints**: the Football-Data.org API imposes request limits, necessitating the implementation of sleep intervals and retry mechanisms.

3.1.4 Importance and Practical Implications

Both processes are indispensable for building **robust**, **reliable** and **actionable** datasets. *Data integration* ensures that no relevant source is overlooked, while *data enrichment* maximizes the informational value of each data point. In combination, they provide a powerful foundation for **insight-driven decision making**, **predictive modeling** and **high-resolution analytics**.

In the context of sports analytics, particularly for the Italian Serie A 2024–2025 season, applying these techniques allows for:

- Merging player statistics and market data into a unified player profile.
- Enriching basic performance statistics with advanced metrics such as Expected Goals (xG), pass value or spatial movement data.
- Enabling multi-layered analysis across technical, physical and tactical domains.

Ultimately, investing effort into data integration and enrichment transforms raw, isolated data into a rich knowledge base capable of supporting complex analytical tasks and delivering strategic value.

4 Data Preparation and Cleaning

The *data cleaning and preparation* phase aimed to harmonize, clean and integrate heterogeneous data sources related to Serie A players. Specifically, we worked with three main types of information: players' market values, seasonal performance statistics and contractual details.

4.1 Standardization of Market Values

Market value data contained heterogeneous formats, with values expressed in millions or thousands of euros (e.g. "€3.5 mln", "€500 mila"). We applied a dedicated cleaning and conversion function to:

- remove currency symbols and non-numeric characters,
- convert all values into integer euro amounts (e.g. €3.5 mln → 3,500,000),
- handle missing or unavailable values (assigning *null*).

This process produced a new standardized numerical column *Market Value (EUR)*, suitable for quantitative analysis.

4.2 Processing Contract Data

The contractual dataset included the contract end date in a textual format (e.g. "2026-06"). From this field, we extracted the contract end year and computed the **remaining contract duration** relative to the reference year 2025 using:

$$\text{Remaining Duration} = \text{Contract End Year} - 2025$$

This new numeric attribute allows us to consider contractual stability as a potential explanatory variable.

4.3 Merging Market Value with Performance Data

For each team, we performed a *merge* operation between the market value dataset and the player performance dataset, using the player name (*Player*) as the join key. A **left join** strategy ensured that all players with performance statistics were retained, with market value information added when available.

We tracked the loss of observations caused by name mismatches or missing market values. For each team, we calculated the percentage of rows lost due to merging issues to document data integration quality.

4.4 Concatenation of All Teams

Once the cleaning and merging were completed for individual teams, all team-level dataframes were vertically concatenated into a single comprehensive Serie A dataset. This operation ensured a consistent column schema across all clubs.

4.5 Imputation of Missing Values

After merging and concatenating the data, we identified missing values in four key columns: *Team*, *Years of Contract Remaining*, *Position* and *Nation*. To ensure data quality and consistency, we applied a systematic imputation strategy, detailed below::

- For the **Team** column, a rule-based approach was used to infer missing values by examining adjacent rows. If a missing entry was surrounded by identical team names, it was filled accordingly. Residual missing values were resolved manually based on domain knowledge about the players' real club affiliations.
- For the **Years of Contract Remaining** column, missing values were filled by consulting public records and transfer information to estimate remaining contract durations. This manual imputation ensured realistic and consistent values across the dataset.
- For the **Position** column, missing values were inferred by analyzing historical match data and player profiles, ensuring that the assigned positions reflected the players' typical roles on the field.
- For the **Nation** column, missing entries were imputed using external databases and official player nationality records, guaranteeing accurate representation of players' countries.

This step produced a complete, clean dataset without missing values in these critical columns, supporting reliable downstream analyses.

4.6 Recoding Object Variables to Numeric

Certain variables in the dataset, such as *Age*, were originally stored as strings (object type) in non-uniform formats that required transformation for quantitative analysis. For example, the *Age* column sometimes contained values like "25-096" (representing an age range). To standardize this variable:

- We split the string on the hyphen character and retained only the first component, interpreting it as the player's minimum age (e.g. "25-096" $\rightarrow$ 25).
- We converted the resulting string values to **float** data type to enable numerical computations.

This step ensured that the *Age* variable was consistently represented as a numeric feature across the entire dataset, supporting robust statistical and predictive modeling.

4.7 Cleaning the `Nation` Column

The `Nation` feature in the dataset initially contains values in the format of a lowercase language code followed by an uppercase country code (e.g. `it ITA`, `pt POR`, `en ENG`). For the purposes of modeling, only the country code is meaningful as a categorical feature representing the player's nationality.

To extract the relevant information, we split each string by whitespace and retain only the uppercase country code.

This ensures that the `Nation` column contains standardized, consistent categorical values (e.g. `ITA`, `POR`, `ENG`) suitable for encoding and inclusion in machine learning pipelines.

4.8 Normalization of Player Positions

The original dataset exhibited considerable heterogeneity in the *Position* column, where players were often assigned multiple roles recorded in varying formats (e.g. "DF-MF", "FW,MF", "MF/FW"). To establish a standardized classification suitable for analysis, we applied the following normalization procedure:

- **Multi-role separation**: Using regular expressions, we identified common delimiters such as hyphens, commas and slashes to split multiple positions into individual roles.
- **Primary position extraction**: For players with multiple listed positions, only the first position was retained, as it generally corresponds to the player's primary role. Examples of the transformation include:
  - "DF-MF" $\rightarrow$ "DF"
  - "FW,MF" $\rightarrow$ "FW"
  - "MF/FW" $\rightarrow$ "MF"
- **Standardization**: All position abbreviations were converted to uppercase to ensure uniformity (e.g. "df" $\rightarrow$ "DF", "mf" $\rightarrow$ "MF").
- **Validation**: The normalized positions were checked against a predefined set of standard football roles {GK, DF, MF, FW} to ensure data validity.

This standardized classification facilitates robust analysis of the relationship between player roles and other variables, such as market value and performance metrics, while reducing complexity in subsequent modeling steps. The approach is consistent with common football analytics practices, where a player's primary position is typically more relevant than secondary roles for most analyses.

4.9 Final Cleaning and Export

Lastly, we removed redundant or unnecessary columns (e.g. residual indices from previous saves) to ensure clarity and consistency. The final dataset was exported as a CSV file, ready for further exploratory analysis or predictive modeling.

4.10 Final Output

The output of the data cleaning and preparation phase is a unified Serie A player dataset containing:

- Individual performance statistics,
- Market value in integer euro amounts,
- Remaining contract duration in years,
- Club identifier.

This clean and integrated dataset serves as the foundation for all subsequent descriptive and predictive analyses in the project.

5 Dataset Description

The dataset created for this project consists of 209 variables describing various aspects of football players' profiles and performances.

To facilitate understanding, the variables have been grouped into thematic categories:

- **Standard Stats**: includes key groups of metrics such as **Playing Time** (e.g. Total minutes, games started by a player, ecc.), **Performance Indicators** (e.g. Goals, assists, ecc.), **Expected Values** (e.g. Expected goals, ecc.), **Progression Metrics** (e.g. Progresses passes, ecc.) and **Per 90 Minutes Statistics**. The essential toolkit to measure playing time, production and proactive impact.
- **Goalkeeping Stats**: features core metrics including **Performance Metrics** (e.g. Goal Against, Saves, ecc.), **Penalty Kicks Indicators** (e.g. Penalty Kicks Allowed, ecc.), **Goals Values** (e.g. Goals Against, Corner Kicks Goals Against, ecc.), **Expected Metrics** (e.g. Post-Shot Expected Goals, ecc.), **Launched** (e.g. Passes Completed (Launched), ecc.), **Passes** (e.g. Passes Attempted, ecc.), **Goal Kicks**, **Corsses** and **Sweeper** (e.g. Defensive Outside Penalty Area, ecc.). A 360° profile covering reflexes, positioning, footwork and anticipation.
- **Shooting**: tracks fundamental performance indicators like **Standard Statistics** (e.g. Shots on Target, Total Shots, ecc.) and **Expected Metrics** (e.g. Non-Penalty Expected Goals, ecc.). These metrics reveal shooting quality and striker's instinct, measuring both volume and true threat of chances.
- **Passing**: captures essential data points such as **Total Statistics** (e.g. Passes Completed, Pass Completion Percentage, ecc.), **Short Passes Values**, **Medium Passes Values** and **Long Passes Values**. This matrix assesses a player's role in possession phases, from secure circulation to chance creation, while revealing their risk/reward balance in different pitch zones.
- **Pass Types**: monitors key performance aspects including **Pass Types Metrics** (e.g. Live-Ball Passes, Through Balls, ecc.), **Corner Kicks** and **Outcome Values** (e.g. Passes Offside, ecc.). This breakdown reveals a player's spatial awareness and technical range, measuring both routine circulation and high-risk/high-reward distributive attempts.
- **Goal and Shot Creation**: analyzes critical metrics covering **Shot Creating Actions (SCA)**, **SCA Types**, **Goal Creating Actions (GCA)** and **GCA Types**. This framework tracks both quantity and methods of chance generation, revealing how players convert possession into high-value opportunities through different creative channels.
- **Defensive Actions**: incorporates primary indicators like **Tackles Statistics** (e.g. Tackles Won, ecc.), **Challenges Metrics** (e.g. Dribblers Tackled, Challenges Lost, ecc.) and **Blocks Metrics** (e.g. Shots Blocked, Passes Blocked, ecc.). These metrics collectively assess a player's ability to break opposition attacks, measuring both proactive interventions and recovery success in defensive situations.
- **Possession**: evaluates principal statistics including **Touches Metrics** (e.g. Touches in defensive, Touches in Middle, ecc.), **Take-Ons Values** (e.g. Succesfull Take-Ons, ecc.), **Carries Metrics** (e.g. Total Carrying Distance, ecc.) and **Receiving** (e.g. Passes Received, ecc.). These metrics collectively map a player's spatial influence and ball retention

effectiveness, revealing how they dictate tempo and create advantages through controlled possession.

- **Playing Time**: documents crucial measurements such as **Starts Metrics** (e.g. Complete Matches Played, ecc.), **Subs Values** (e.g. Substitute Appearances, ecc.), **Team Success Indicators** (e.g. Points per Match, Goal Scored (While on pitch), ecc.) and **Team Success (Expected Goals)**. This comprehensive framework quantifies both raw participation and contextual effectiveness, revealing how players influence games whether starting or coming off the bench.
- **Miscellaneous Stats**: key performance metrics, including **Performances Statistics** (e.g. Yellow Cards, Offsides, ecc.) and **Aerials Duels** (e.g. Aerials Won, Aerials Lost, ecc.). These multifaceted metrics capture the nuanced aspects of performance, from disciplinary actions to aerial dominance, painting a complete picture beyond conventional stats.
- **Market Value (EUR)**: market value (EUR) reflects the ever-shifting balance of demand, potential and perceived worth in a dynamic financial landscape.
- **Years of Contract Remaining**[33]: the remaining years on a contract define the balance between commitment and flexibility, shaping decisions in sports, business and beyond.

For a complete description of the 209 variables used in this analysis, please refer to the attached file available at the following link:

Variables Description

6 Data Quality

The final database consists of 545 rows and 209 columns, with each row representing a Serie A player for the 2024–2025 season. The columns include detailed performance metrics (e.g. Tackles, Carries, xG), contractual information (Years of Contract Remaining), team affiliation and Market Value (EUR).

6.1 Completeness

Completeness refers to the extent to which all required data is present in the dataset. Incomplete data can introduce bias or negatively affect the accuracy of analyses. Therefore, it is crucial to identify missing values and, when possible, fill them using reliable imputation methods or manual recovery from trusted sources. In our Database we have some statistics that may be absent due to player's role (e.g. , attackers do not save penalties).

During the data integration process, the team identified missing values in several key variables. The following table summarizes the completeness analysis:

Variable	**Missing Values (%)**
Nation	14.7%
Pos	14.5%
Years of Contract Remaining	22.22%
Team	17.8%
Market Value	∼38%

Table 1: Summary of missing values

The gaps primarily resulted from discrepancies between web-scraped data and API sources. For most variables, manual imputations were performed using reliable sources (e.g. official squad lists and player profiles) to ensure critical fields weren't empty. The final dataset shows 62.17% completeness, with the main incompleteness coming from market values (missing for many younger players via Transfermarkt).

6.2 Consistency

Consistency indicates how coherent data is across different sources and within the same dataset. Inconsistencies can arise from differing formats, typographical errors or varying coding standards. Normalizing data and applying uniform rules reduce these discrepancies and make data integrable and comparable.

Data consistency was a major focus during the cleaning phase. Player names often exhibited variations in spelling or formatting across different sources. The team applied a systematic normalization process, standardizing letter casing and removing unnecessary characters to enable precise merges. Additionally, consistent formatting was enforced across all categorical and numerical fields to facilitate robust analysis.

6.3 Accuracy

Accuracy relates to how well the data represents the true values or reality. Inaccurate data may result from collection errors, imprecise merging or incorrect imputations. Ensuring accuracy involves cross-checking data against trusted sources and using integration procedures that maintain correctness.

To improve data accuracy, thorough verification steps were implemented. During merging, left joins were carefully applied to retain the primary dataset's structure while excluding non-matching or conflicting records. The team also cross-validated manually imputed fields against trustworthy sources to minimize errors.

6.4 Timeliness

Timeliness concerns the relevance of data with respect to time. Outdated or stale data can reduce reliability, especially in dynamic contexts such as sports or markets. It is important to filter obsolete or duplicated records and maintain only the most recent information to reflect the current situation accurately.

Records with outdated or duplicate entries were identified and removed. For players appearing multiple times, only the version with the most recent statistics was retained. This ensured that the final dataset accurately reflected the 2024–2025 season.

6.5 Validity

Validity refers to conformity of data to predefined rules and constraints, such as correct formats and plausible values. Checking data types and acceptable ranges helps detect errors—e.g. out-of-range numeric values or malformed text fields—and ensures the data is suitable for quantitative and qualitative analysis.

Data types were systematically checked and corrected where necessary. For example, columns such as *Age* were initially encoded as strings (objects) due to inconsistent formats, and were then converted then recoded as numeric (float) types to support quantitative analyses. We also conducted range checks on variables (e.g. verifying plausible age values and ensuring percentages were within 0–100) to uphold validity rules.

6.6 Uniqueness

Uniqueness ensures that each entity in the dataset is represented only once, avoiding duplicates that can distort results. Detecting and removing duplicate records is fundamental to preserving dataset integrity and ensuring analyses are based on independent observations.

Duplicate records were detected and eliminated to guarantee that each player appeared only once in the final dataset. The criteria for deduplication prioritized the inclusion of the most complete and up-to-date observations.

All cleaning and integration tasks were performed using Python, primarily leveraging the `pandas` library for data manipulation, merging, imputation and quality checks.

7 Data Storage

Choosing an appropriate data storage solution is a critical decision in data science, as it directly impacts data accessibility, scalability and analytical efficiency. The optimal choice depends on factors such as data volume, structure, query complexity and performance requirements. Below, we discuss the primary storage options available.

7.1 Relational Databases (SQL-Based)

Relational databases organize data into structured tables with predefined schemas and relationships.

- **Examples:** MySQL, SQLite.
- **Pros:**
  - Structured schema enforcement ensures data consistency.
  - Powerful querying capabilities (JOINs, aggregations) via SQL.
  - ACID (Atomicity, Consistency, Isolation, Durability) compliance for transactional reliability.
- **Cons:**
  - Vertical scaling can be expensive.
  - Less flexible for unstructured or rapidly evolving data.
- **Best for:** Structured data with complex relationships (e.g. transactional systems).

7.2 NoSQL Databases

NoSQL databases provide schema flexibility and are designed for scalability and performance in distributed environments.

7.2.1 Document Stores (MongoDB)

- Store data in JSON-like documents.
- Ideal for semi-structured or evolving data.

7.2.2 Key-Value Stores (Redis, DynamoDB)

- Optimized for high-speed read/write operations.
- Commonly used for caching and session storage.

7.2.3 Columnar Databases (Cassandra, HBase)

- Optimized for large-scale analytics and time-series data.
- Efficient for queries over large datasets.

7.2.4 Graph Databases (Neo4j)

- Designed for data with complex relationships (e.g. social networks).
- Use graph structures for semantic queries.
- **Pros:**
  - Horizontal scalability.
  - Flexible schema design.
  - High performance for specific use cases.
- **Cons:**
  - Lack of standardized querying (varies by database).
  - Eventual consistency trade-offs in distributed systems.
- **Best for:** Unstructured/semi-structured data, real-time applications or distributed systems.

7.3 Data Warehouses

Data warehouses are optimized for analytical processing and large-scale data storage.

- **Examples:** Snowflake, Google BigQuery, Amazon Redshift.
- **Pros:**
  - Optimized for OLAP (Online Analytical Processing).
  - Columnar storage improves query performance.
  - Parallel query execution for large datasets.
- **Cons:**
  - Higher cost compared to traditional databases.
  - Less efficient for transactional workloads (OLTP).
- **Best for:** Large-scale analytics and business intelligence.

7.4 Distributed File Systems & Big Data Tools

These systems are designed for handling massive datasets across distributed clusters.

- **Examples:** Hadoop HDFS, Apache Parquet, Delta Lake.
- **Pros:**
  - Fault-tolerant and scalable.
  - Cost-effective for petabyte-scale storage.
  - Compatible with distributed processing frameworks (e.g. Apache Spark).
- **Cons:**

  - Steeper learning curve.
  - Overhead for small datasets.
- **Best for:** Batch processing, data lakes and ETL pipelines.

7.5 In-Memory Databases

In-memory databases store data primarily in RAM for ultra-fast access.

- **Examples:** Redis, Apache Ignite.
- **Pros:**
  - Sub-millisecond latency.
  - Ideal for real-time applications.
- **Cons:**
  - Volatile storage (unless persisted to disk).
  - High RAM requirements can be expensive.
- **Best for:** Caching, session storage or high-frequency trading systems.

7.6 Key Considerations for Selection

When choosing a data storage solution, we have to consider the following factors:

- **Data Structure**: structured data fits well with SQL, while unstructured data is better handled by NoSQL.
- **Scalability**: vertical scaling (SQL) vs. horizontal scaling (NoSQL/Data Warehouses).
- **Access Patterns**: frequent writes may favor Key-Value stores, whereas complex analytics benefit from Columnar databases.
- **Budget**: cloud-based solutions (e.g. BigQuery) offer convenience but may incur higher costs compared to self-hosted options (e.g. PostgreSQL).

7.7 Why SQL?

SQL (Structured Query Language) was chosen as the foundation for managing the Serie A players dataset because its design and capabilities align precisely with the goals of this project: supporting player market research questions and enabling downstream machine learning (ML) modeling. While NoSQL systems can be beneficial for unstructured or highly variable data, this project required a structured, relational approach to ensure data integrity, ease of querying and analytical power.

7.7.1 Purpose-Driven Database Design

The database was deliberately structured as a single, well-defined table where:

- Each **row** represents an individual Serie A player.

- Each **column** captures a specific variable, including performance metrics, contractual details and standardized market value in euros.

This schema was specifically engineered to answer market-oriented research questions such as:

- Which performance indicators correlate most strongly with market value?
- How does remaining contract duration impact player valuation?
- How do clubs differ in their player profiles and value distribution?

By storing the data in a single, relationally consistent table, SQL enables these questions to be answered clearly and reproducibly.

7.7.2 Intentional Support for Machine Learning Workflows

The SQL database also facilitates the transition from raw data storage to ML-ready datasets. The table structure supports:

- Easy extraction of features and labels for predictive modeling.
- Straightforward filtering, joining and aggregation to generate training sets.
- Consistent data types and integrity constraints that reduce preprocessing effort.

Choosing SQL was therefore not only about data storage, but about aligning the data model with the intended use case of developing predictive models for player valuation.

7.7.3 Declarative and Expressive Querying

SQL's declarative syntax allows researchers to specify *what* they want from the data without defining *how* to retrieve it. This is essential for:

- Writing clear, maintainable queries to select, filter and transform variables.
- Rapidly prototyping new research questions without reengineering data pipelines.

Features like Common Table Expressions (CTEs), window functions and complex joins enable highly expressive analytical workflows directly within the database.

7.7.4 Data Integrity and Reliability

Market research and ML depend on trustworthy data. SQL databases enforce ACID properties:

- **Atomicity**: ensures all-or-nothing updates for critical cleaning steps.
- **Consistency**: enforces valid relationships and data types.
- **Isolation**: supports concurrent access without corruption.
- **Durability:** guarantees persistence even in case of failures.

This robustness is particularly valuable when integrating heterogeneous sources such as market values, performance statistics and contract data.

7.7.5 Scalability for Research and Analysis

Although the current database is a single table of Serie A players, SQL systems can easily scale to include:

- Historical seasons for longitudinal analysis.
- Multiple leagues for comparative research.
- Related tables for transfer history or injury records.

This future-proof design ensures the database remains useful as research questions expand.

SQL was chosen deliberately to match the data's structured nature and the project's analytical goals. Its ability to ensure data integrity, enable expressive research queries and support ML workflows makes it the ideal choice for managing a clean, unified dataset of Serie A players designed to answer real-world market valuation questions.

7.8 Rationale for Adopting a Unified Table Structure

In developing a predictive model for the market value of Serie A football players, we intentionally designed the database to store all player-related features—regardless of role—in a **single relational table**, using `NULL` values for statistics not applicable to certain roles. This unified schema was chosen for several technical and practical reasons:

1. **Semantic Integrity of Missing Values**
   Many player statistics are inherently role-dependent. For example:
   - **Goalkeepers** have metrics such as `saves`, `clean sheets` and `claims`, which are **irrelevant** for outfield players.
   - **Outfield players** have statistics like `goals`, `assists` or `dribbles`, which do not apply to goalkeepers.

   Using `NULL` in these cases preserves **semantic clarity**: it explicitly indicates "not applicable," avoiding the ambiguity of encoding these as `0`, which could incorrectly suggest "poor performance."

2. **Native Compatibility with Tree-Based Machine Learning Algorithms**
   Modern gradient boosting frameworks such as **CatBoost**, **XGBoost** and **LightGBM** natively support `NULL` (or `NaN`) values:
   - They **learn optimal default splits** for missing values during training.
   - This capability eliminates the need for potentially biased imputation strategies and allows the model to **distinguish between a true zero and a non-applicable field**.

3. **Simplified Data Management and Feature Engineering**
   Maintaining a **single table** schema enables:
   - **Uniform access** to all data, regardless of player role.
   - Easier creation of **derived features**.
   - Streamlined batch transformations and scaling within ML pipelines.

By contrast, using multiple role-specific tables would require complex joins, custom preprocessing pipelines per role and could introduce redundant logic.

4. **Improved Model Interpretability and Flexibility**
   Retaining all features in a unified schema with NULL for non-applicable fields enables training a single, general-purpose model that is simpler to interpret and easier to maintain in production.

5. **Mitigation of Data Leakage Risk**
   Encoding non-applicable statistics as `0` can introduce **spurious patterns** during model training:

   - For example, attackers would systematically show `0 saves`, potentially leading the model to misinterpret this as poor goalkeeping rather than "not applicable."

   Proper use of `NULL` ensures the model does not **overfit to role-specific artifacts** introduced by incorrect imputation.

Adopting a **single table with `NULL` values for non-applicable statistics** ensures semantic correctness, leverages the native capabilities of modern ML algorithms, simplifies data engineering workflows and improves model interpretability. This approach aligns with both **relational design best practices** and **modern machine learning standards**.

8 Exploratory Data Analyisis (EDA)

Exploratory Data Analysis (EDA) is a fundamental step in the data analysis workflow. Its primary goal is to **understand the structure, characteristics and relationships within the data** before applying predictive or inferential models.

During this phase, analysts typically focus on:

- Examining the **distribution of variables**;
- Identifying **outliers** and unusual patterns;
- Formulating **preliminary hypotheses** and generating insights.

EDA is crucial because it allows for a **deep understanding of the dataset** prior to the application of more complex models. It helps prevent misinterpretations, improve data quality and design more targeted analytical strategies.

The following plots illustrate only a small selection of the many possible insights that can be derived from the Serie A dataset. Given the dataset's extent and complexity, it is not feasible to include every potential visualization here. Instead, these examples were chosen to highlight some of the most relevant trends and patterns, while acknowledging that further exploration could reveal many additional insights.

8.1 General Overview of Serie A 2024-25

As a general introduction to the 2024/25 Serie A season, the graphic [1] provides a visual representation of the **total market value of each team**, measured in **millions of euros**. This snapshot highlights the **economic disparities** that exist across clubs in the league.

At the top end of the distribution, teams such as *FC Internazionale Milano*, *Juventus FC* and *AC Milan* stand out, each exceeding **€500 million** in total squad value. These figures reflect not only their **financial capacity** but also the depth and quality of their player rosters.

Conversely, clubs like *Venezia*, *US Lecce* and *Monza* occupy the lower end of the spectrum, with significantly more limited market values. These differences suggest the presence of a **dual-speed league**, where resource availability directly impacts competitiveness.

This economic gap has concrete implications for Serie A's competitive landscape, influencing areas such as **transfer strategy**, **youth development** and overall team performance. Understanding this financial stratification is essential to contextualize the subsequent analyses of player attributes and team dynamics throughout the season.

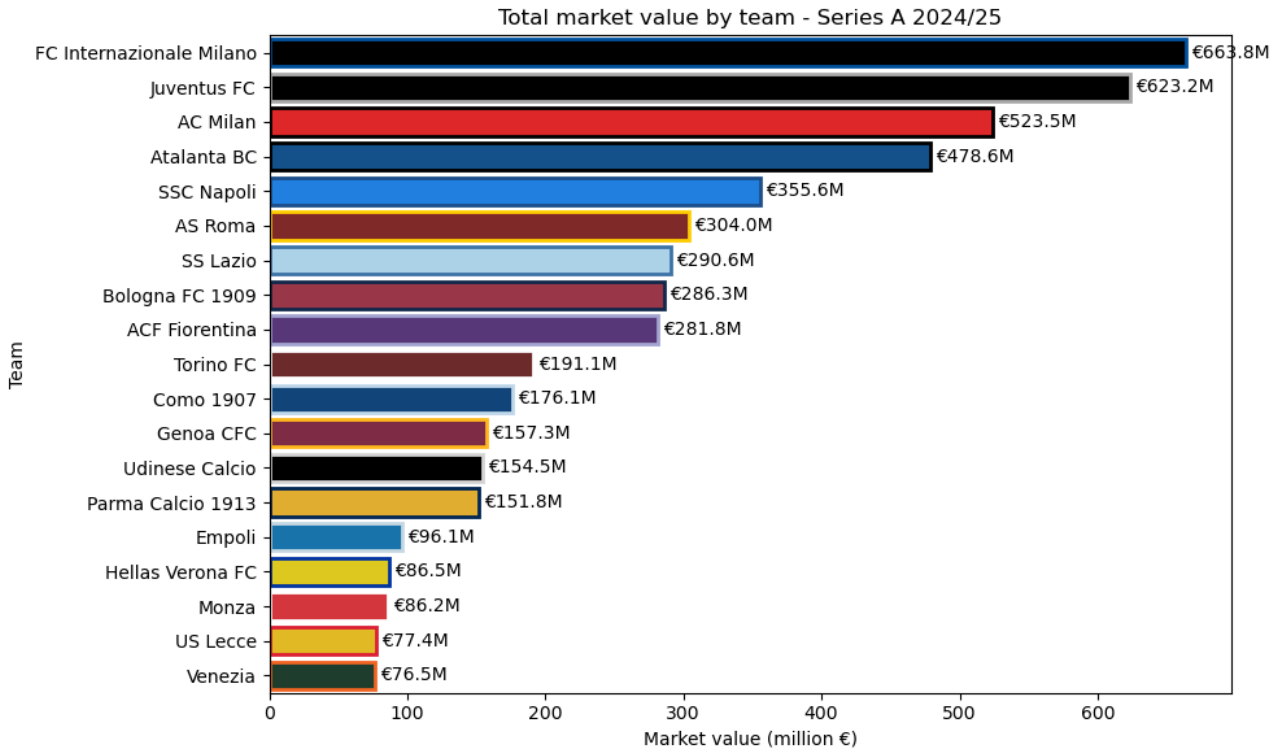

Figure 1: Total Players market value by team.

The plot [2] presents the **average age of players** for each Serie A team in the 2024/25 season. This metric offers valuable insight into the **balance between youth and experience** within team rosters, which can shape a club's **tactical identity**, **playing style** and **long-term development strategy**.

At one end of the spectrum, *FC Internazionale Milano* has the **highest average age**, nearing **29.5 years**, indicating a veteran squad likely designed to maximize short-term success in both domestic and international competitions. Such a profile often suggests an emphasis on **immediate competitiveness**, supported by seasoned professionals.

Conversely, clubs like *Parma Calcio 1913*, *Juventus FC* and *US Lecce* exhibit significantly **younger average ages**. These profiles may reflect different strategic orientations: a commitment to **youth development**, the integration of academy talent or the need to operate within **tighter financial constraints**, limiting reliance on costly veteran acquisitions.

Analyzing the age composition of each squad is crucial for interpreting performance patterns throughout the season. Teams with younger rosters might demonstrate greater physical dynamism and growth potential, while more experienced squads could rely on tactical discipline and game management, particularly in high-stakes matches.

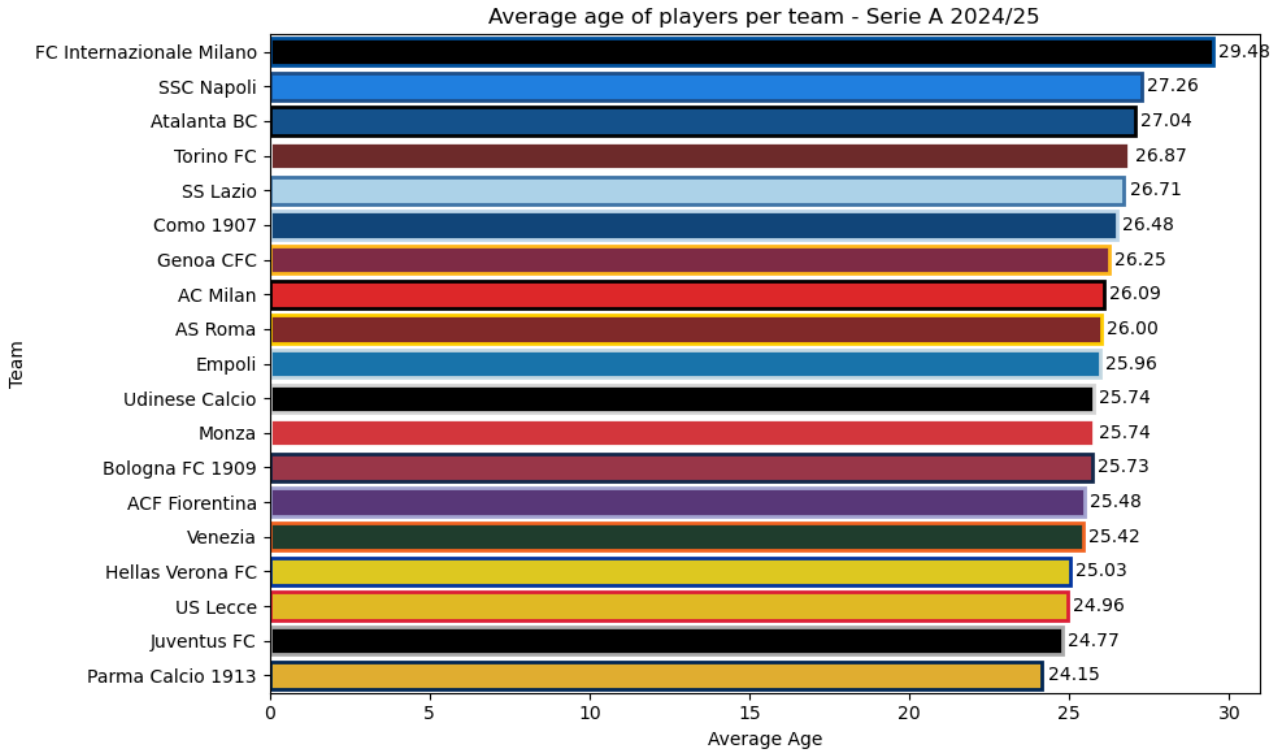

Figure 2: Average Players Age by team.

8.2 Players Analysis

Figure [3] presents a **side-by-side comparison** between the top 10 players in terms of **Expected Goals (xG)** and their corresponding **actual goal counts** for the 2024/25 Serie A season. This visualization offers critical insight into each player's **scoring efficiency** and the degree to which they **outperform or underperform** relative to the quality of chances they accumulate.

Mateo Retegui emerges as a standout performer, topping both rankings with an xG of **19.0** and scoring an impressive **25 goals**. This indicates a significant level of **overperformance**, suggesting clinical finishing or an ability to convert difficult chances. Similarly, players like *Moise Kean* and *Artem Dovbyk* exceeded their expected returns, reinforcing their reputations as **efficient and effective finishers**.

In contrast, players such as *Romelu Lukaku* and *Nikola Krstovic* underperformed relative to their xG, implying possible **finishing inefficiencies**, poor shot placement or variance in match situations. Such discrepancies between xG and actual goals can also reflect differences in **tactical roles**, **confidence levels** or even **luck** over the course of the season.

Overall, this dual-panel graph provides a valuable lens through which to assess **individual attacking contributions**, differentiating between players who thrive on high xG volumes and those who maximize fewer opportunities through exceptional execution.

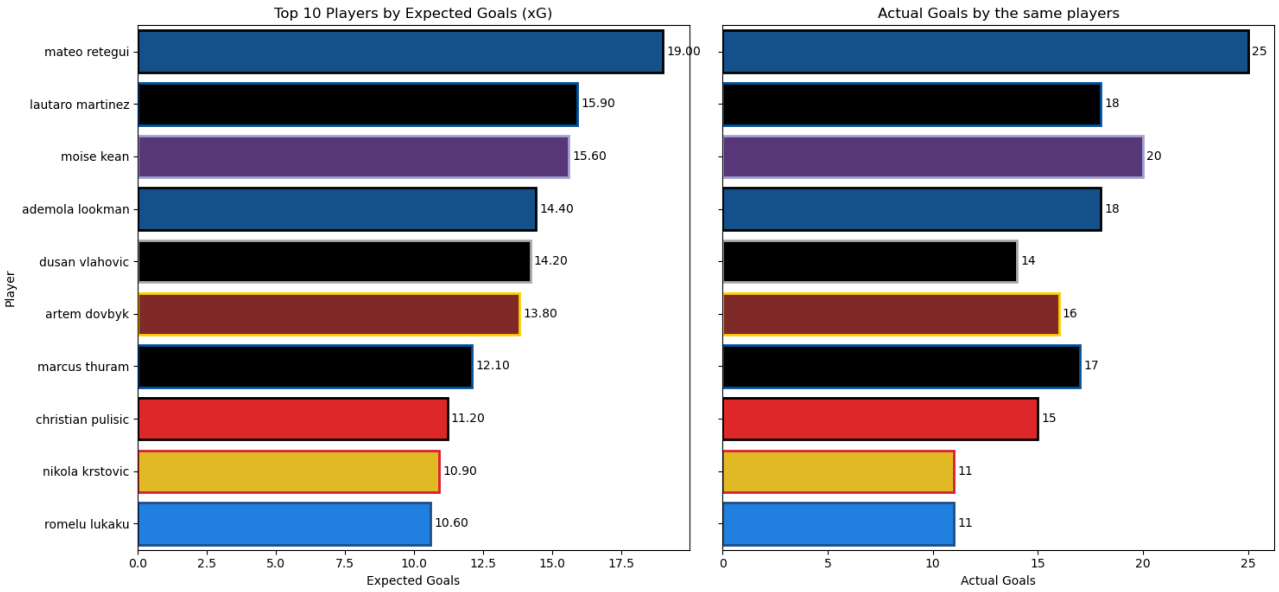

Figure 3: Expected Goals vs Actual Goals.

The boxplot [4] illustrates the distribution of player market values (in millions of euros) across four primary roles in Serie A for the 2024/25 season: **Goalkeepers (GK)**, **Defenders (DF)**, **Midfielders (MF)** and **Forwards (FW)**.

As expected, **forwards** show the **highest median** and **widest spread** in market values, underlining their crucial importance in the attacking phase. The presence of several high-value outliers reflects a market that **strongly rewards offensive talent**.

Midfielders also display a **broad value distribution**, with many players reaching or even exceeding the valuations of defenders and goalkeepers. This underlines their hybrid role in linking defense and attack, and their strategic importance in modern football.

Defenders exhibit a **moderate median value** with a fairly wide interquartile range, suggesting a variety of profiles and market recognition within this role. The presence of several outliers indicates that **top-level defenders remain highly valued assets**.

In contrast, **goalkeepers** show the **lowest median market value** and the most compact distribution, reflecting a more standardized valuation in this role. Nonetheless, a few exceptional goalkeepers still reach significantly higher values, as evidenced by the outliers.

Overall, the chart highlights the **market hierarchy by position** in Serie A: attacking roles (FW, MF) dominate in terms of value, while defensive and goalkeeping roles show more consistent but generally lower valuations. This reflects the broader trend in football economics, where creative and goal-scoring abilities command a premium.

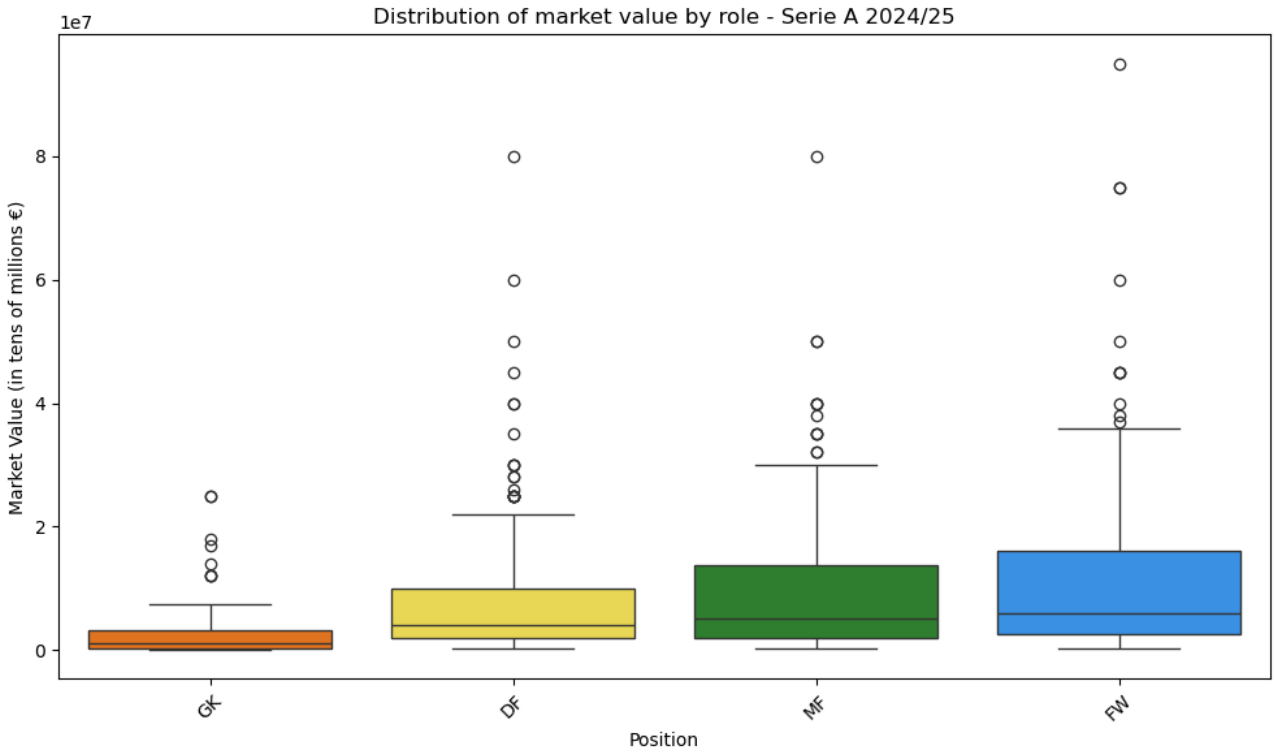

Figure 4: Players Market Value Distribution by Role.

9 Football Players Market Value Prediction

9.1 Problem Definition

The objective of this section is to predict the *Market Value (EUR)* of football players in the Serie A for the 2024–25 season. Accurate prediction of market value can provide valuable insights for scouting, negotiations and transfer strategies, helping clubs identify potentially overvalued or undervalued players. For example, detecting systematic overvaluation can help avoid costly transfers, while spotting undervalued players can lead to smart acquisitions that strengthen the squad at a lower cost. The target variable used in this analysis is the players' reported market value in euros.

9.2 Model Selection

After evaluating different regression algorithms, we chose to use the CatBoostRegressor for this task. This model offers key advantages when dealing with real-world football data, as it automatically learns optimal default splits for missing values during training. This feature eliminates the need for potentially biased imputation strategies and enables the model to distinguish between true zeros and genuinely non-applicable fields.

These properties make CatBoost particularly well-suited for football data, which often includes missing or context-dependent features, such as player appearances or goals.ndent features such as appearances or goals.

9.3 Model Training and Evaluation

To evaluate model performance rigorously while avoiding overfitting, we adopted a **Leave-One-Out Cross-Validation (LOOCV)** strategy. In this approach, each player in the dataset is left out once as a test instance, while the model is trained on all remaining players. This procedure ensures maximum use of available data and provides an unbiased estimate of prediction error, especially valuable given the relatively small sample size.

We applied a *logarithmic transformation* to the target variable (market value) to stabilize variance and reduce the impact of extreme values. Additionally, we performed outlier filtering on the training subset for each fold by excluding the top and bottom 1% of log-transformed values, thereby mitigating undue influence from anomalous transfer fees.

The loss function used was *RMSE* in log space, while predictions were converted back to the original scale via exponentiation.

To summarize prediction accuracy, we computed the **Mean Absolute Error (MAE)** across all LOOCV folds. The model achieved an MAE of approximately *€ 4,196,636* on the hold-out players. This result indicates that, on average, the predicted market value differs from the actual value by roughly this amount per player—a level of error that is reasonable given the inherent noise and subjectivity in transfer market valuations.

9.4 Results and Discussion

The model's predictions offer a data-driven perspective on player valuation, which can support strategic decision-making in the transfer market. However, a key pattern emerged from the results: the model tends to **underestimate the market value of high-profile players** and **overestimate the value of lower-ranked or lesser-known players**.

This behavior may stem from the fact that extreme market valuations are often influenced by non-observable or intangible factors—such as media attention, commercial appeal or unique contractual terms—that are difficult to capture with standard performance and demographic features. As a result, the model regresses toward the mean, producing conservative estimates for outliers.

For instance, if a player like Leão has a known market value of €75 million but the model predicts only €33 million, it suggests a potential overvaluation from a purely data-driven standpoint—or alternatively, a limitation of the model in capturing intangible market drivers. Conversely, when undervalued players receive higher predicted prices, it may highlight interesting acquisition opportunities for clubs willing to trust in objective metrics.

A separate feature importance analysis (see next section) reveals which variables most strongly influence predicted value, offering further insight into the components driving model predictions and how clubs might act on them.

Overall, while the model provides useful approximations, its limitations—particularly in modeling outliers—should be considered when interpreting predictions in the context of real-world transfer dynamics.

9.5 Feature Importance Analysis

To further interpret the model's predictions, we analyzed the relative importance of each input variable using CatBoost's built-in feature importance metric. Figure 5 shows the top 20 features ranked by their contribution to reducing prediction error.

The most influential features were:

- **Age**: younger players tend to command higher fees due to greater potential resale value and career longevity.
- **Team**: captures differences in exposure, reputationand club resources that affect player valuation.
- **Years of Contract Remaining**: longer contracts typically increase transfer costs by giving clubs greater negotiating power.
- **TeamSuccess_onG**, **TeamSuccess_xG_onxG**, **TeamSuccess_+/-**: metrics summarizing team-level success and performance that can elevate a player's market value.
- **Subs_unSub**, **GCA**, **SCA**: playing time, creative contributions and substitutions that indicate the player's role and coach trust.

These results suggest that both individual characteristics (age, contract details, on-ball actions) and team context (club, team success metrics) are critical drivers of market value predictions. Notably, features describing team-level performance (expected goals, goal difference)

were consistently important, supporting the idea that player valuation depends partly on team environment and results.

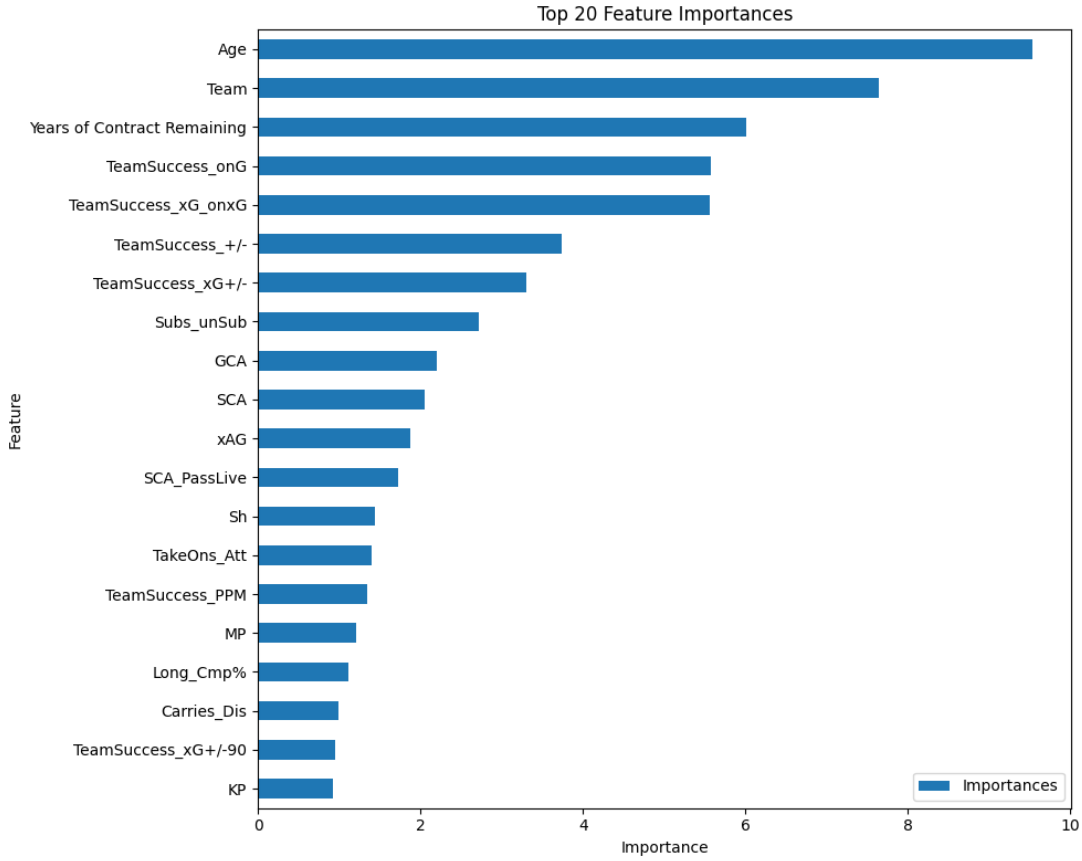

Figure 5: Top 20 feature importances derived from the CatBoostRegressor model.

It is important to note that these rankings reflect the available dataset and modeling approach. Factors like international reputation, commercial appeal or special contractual clauses, which can strongly affect real-world valuations, are not captured in this analysis. Therefore, feature importance results should be interpreted as complementary to traditional scouting and negotiation expertise.

9.6 Approach and Results

We developed and evaluated a machine learning model to predict football players' market value in the Serie A 2024–25 season. Using `CatBoostRegressor`, the model effectively handled missing data and delivered competitive predictive performance.

Such a tool can help clubs identify pricing inefficiencies, supporting data-driven transfer strategies. Future improvements could include enriching the dataset with additional features (e.g. injury history, social media metrics) and exploring ensemble or neural network approaches for even better accuracy.

10 Conclusions

This project successfully achieved its primary objective of predicting the market value of Serie A football players for the 2024-2025 season by leveraging their performance statistics. A central component of this effort was the design and implementation of a comprehensive and rigorous data management pipeline, encompassing the phases of data acquisition, integration, cleaning and storage. This pipeline provided a solid foundation that ensured the overall quality, consistency and reliability of the dataset, which was essential for subsequent analytical and predictive tasks.

The data acquisition phase employed a hybrid approach, combining web scraping techniques—using tools such as BeautifulSoup for static HTML content and Selenium for dynamic web pages—from reputable platforms like Transfermarkt and FBref.com. This was complemented by the integration of API data from football-data.org, which provided contractual informations. This multi-source, multi-method strategy allowed the compilation of a rich, scalable and multidimensional dataset capturing both technical performance metrics and economic factors.

Following data collection, a significant effort was dedicated to the integration and enrichment of heterogeneous data sources into a unified, coherent dataset. This process involved applying a consistent player identifier system based on name normalization and utilizing outer join operations to avoid loss of valuable information. Addressing common challenges such as name disambiguation and incomplete data availability—especially for younger or less established players—was crucial to improve data completeness and depth. The resulting dataset effectively combined on-field performance data with economic valuations and contractual details, enabling a comprehensive multidimensional analysis.

The data cleaning and preparation phase was critical to ensure the dataset's usability and accuracy. Key tasks included standardizing market values to integer euro amounts, calculating remaining contract years from raw contract data, merging economic and performance datasets, and concatenating team-level data into a single complete database. Special attention was paid to imputing missing values in crucial fields such as team affiliation, contract length, player position and nationality, using a combination of public records consultation. Additionally, categorical variables were recoded into numerical formats and player positions were normalized to a standardized classification scheme to handle role multiplicity consistently.

Throughout the pipeline, continuous assessment and improvement of data quality were prioritized by focusing on multiple dimensions including completeness, consistency, accuracy, timeliness, validity and uniqueness. Although overall dataset completeness was somewhat limited (approximately 62%) due to missing market values for some players, extensive mitigation efforts through targeted imputation and cross-validation helped minimize potential biases.

For data storage, a relational SQL database was chosen to manage structured data efficiently. This solution ensured data integrity via ACID properties and provided native support for machine learning workflows. The adoption of a unified table structure, allowing NULL values for inapplicable statistics, was instrumental in maintaining semantic clarity and compatibility with tree-based machine learning algorithms such as CatBoost. This design simplified data handling and reduced the need for complex imputations that might otherwise introduce modeling bias.

Finally, exploratory data analysis (EDA) offered valuable insights into the dataset's internal structure and relationships. Noteworthy findings included the economic disparities among Serie A clubs in terms of total team market values, variations in average player age by team,

differences between expected goals (xG) and actual goals among top scorers and the distribution of market values by player position, highlighting the premium placed on offensive roles. These insights were pivotal in guiding feature selection and informing the design of predictive models.

In conclusion, the robustness and completeness of the data management pipeline—from initial data acquisition through to preparation and storage—proved essential for constructing a reliable and comprehensive dataset. This foundation enabled detailed analysis of the Serie A player market landscape for the 2024-2025 season and laid the groundwork for developing predictive models capable of supporting data-driven decision-making in the football transfer market.

11 Future Developments

Building on the foundation established in this project, several avenues exist for further improving both the data pipeline and the predictive modeling approach.

First, enriching the dataset with additional, high-impact features could substantially enhance predictive accuracy. Potential new data sources include detailed injury histories, player fitness metrics, social media engagement indicators and agent or contract negotiation details. Integrating more granular match-level statistics (e.g. progressive passes, defensive duels, xA) could also improve the model's ability to capture nuanced performance differences between players.

Second, expanding data coverage to include other leagues beyond Serie A would allow for the development of more generalizable and transferable models. This would also enable transfer market analysis across leagues, helping clubs identify undervalued players abroad.

From a modeling perspective, exploring advanced ensemble techniques (e.g. stacking, blending) or deep learning approaches (e.g. feedforward or graph neural networks) could improve predictive performance by capturing complex, non-linear relationships in the data. Further, leveraging probabilistic models or quantile regression could provide richer insights by predicting confidence intervals or ranges for player market values rather than single point estimates.

In terms of data management, implementing automated data collection pipelines with robust error handling and monitoring would improve scalability and maintainability. Regular, incremental updates to the dataset would also support real-time or near-real-time market value prediction.

References

[1] Serie A. Atalanta players seasonal stats, 2024-25. https://fbref.com/en/squads/922493f3/2024-2025/all_comps/Atalanta-Stats-All-Competitions.

[2] Serie A. Atalanta's players market value, 2024-25. https://www.transfermarkt.it/atalanta-bergamo/startseite/verein/800/saison_id/2024.

[3] Serie A. Bologna players seasonal stats, 2024-25. https://fbref.com/en/squads/1d8099f8/Bologna-Statss.

[4] Serie A. Bologna's players market value, 2024-25. https://www.transfermarkt.it/bologna-fc/startseite/verein/1025.

[5] Serie A. Cagliari players seasonal stats, 2024-25. https://fbref.com/en/squads/c4260e09/Cagliari-Stats.

[6] Serie A. Cagliari's players market value, 2024-25. https://www.transfermarkt.it/cagliari-calcio/startseite/verein/1390.

[7] Serie A. Como players seasonal stats, 2024-25. https://fbref.com/en/squads/28c9c3cd/Como-Stats.

[8] Serie A. Como's players market value, 2024-25. https://www.transfermarkt.it/como-1907/startseite/verein/1047.

[9] Serie A. Empoli players seasonal stats, 2024-25. https://fbref.com/en/squads/a3d88bd8/Empoli-Stats.

[10] Serie A. Empoli's players market value, 2024-25. https://www.transfermarkt.it/fc-empoli/startseite/verein/749.

[11] Serie A. Fiorentina players seasonal stats, 2024-25. https://fbref.com/en/squads/421387cf/Fiorentina-Stats.

[12] Serie A. Fiorentina's players market value, 2024-25. https://www.transfermarkt.it/ac-florenz/startseite/verein/430.

[13] Serie A. Genoa players seasonal stats, 2024-25. https://fbref.com/en/squads/658bf2de/Genoa-Stats.

[14] Serie A. Genoa's players market value, 2024-25. https://www.transfermarkt.it/genua-cfc/startseite/verein/2520.

[15] Serie A. Hellas verona players seasonal stats, 2024-25. https://fbref.com/en/squads/0e72edf2/Hellas-Verona-Stats.

[16] Serie A. Hellas verona's players market value, 2024-25. https://www.transfermarkt.it/hellas-verona/startseite/verein/276.

[17] Serie A. Inter players seasonal stats, 2024-25. https://fbref.com/en/squads/d609edc0/2024-2025/Internazionale-Stats.

[18] Serie A. Inter's players market value, 2024-25. https://www.transfermarkt.it/inter-mailand/startseite/verein/46.

[19] Serie A. Juventus players seasonal stats, 2024-25. https://fbref.com/en/squads/e0652b02/2024-2025/Juventus-Stats.

[20] Serie A. Juventus's players market value, 2024-25. https://www.transfermarkt.it/juventus-turin/startseite/verein/506.

[21] Serie A. Lazio players seasonal stats, 2024-25. https://fbref.com/en/squads/7213da33/Lazio-Stats.

[22] Serie A. Lazio's players market value, 2024-25. https://www.transfermarkt.it/lazio-rom/startseite/verein/398.

[23] Serie A. Lecce players seasonal stats, 2024-25. https://www.transfermarkt.it/us-lecce/startseite/verein/1005.

[24] Serie A. Lecce's players market value, 2024-25. https://www.transfermarkt.it/us-lecce/startseite/verein/1005.

[25] Serie A. Milan players seasonal stats, 2024-25. https://fbref.com/en/squads/dc56fe14/Milan-Stats.

[26] Serie A. Milan's players market value, 2024-25. https://www.transfermarkt.it/ac-mailand/startseite/verein/5.

[27] Serie A. Monza players seasonal stats, 2024-25. https://fbref.com/en/squads/21680aa4/Monza-Stats.

[28] Serie A. Monza's players market value, 2024-25. https://www.transfermarkt.it/ac-monza/startseite/verein/2919.

[29] Serie A. Napoli players seasonal stats, 2024-25. https://fbref.com/en/squads/d48ad4ff/Napoli-Stats.

[30] Serie A. Napoli's players market value, 2024-25. https://www.transfermarkt.it/ssc-neapel/startseite/verein/6195.

[31] Serie A. Parma players seasonal stats, 2024-25. https://fbref.com/en/squads/eab4234c/Parma-Stats.

[32] Serie A. Parma's players market value, 2024-25. https://www.transfermarkt.it/parma-calcio-1913/startseite/verein/130.

[33] Serie A. Player's contract length, 2024-25. https://api.football-data.org.

[34] Serie A. Roma players seasonal stats, 2024-25. https://fbref.com/en/squads/cf74a709/Roma-Stats.

[35] Serie A. Roma's players market value, 2024-25. https://www.transfermarkt.it/as-rom/startseite/verein/12.

[36] Serie A. Torino players seasonal stats, 2024-25. https://fbref.com/en/squads/105360fe/Torino-Stats.

[37] Serie A. Torino's players market value, 2024-25. https://www.transfermarkt.it/fc-turin/startseite/verein/4162.

[38] Serie A. Udinese players seasonal stats, 2024-25. https://fbref.com/en/squads/04eea015/Udinese-Stats.

[39] Serie A. Udinese's players market value, 2024-25. https://www.transfermarkt.it/udinese-calcio/startseite/verein/410.

[40] Serie A. Venezia players seasonal stats, 2024-25. https://fbref.com/en/squads/04eea015/Udinese-Stats.

[41] Serie A. Venezia's players market value, 2024-25. https://www.transfermarkt.it/venezia-fc/startseite/verein/607.