

Determinants of Soccer Players' Market Values - Italy*

John Zhang

December 3, 2024

1 Motivation

1. **For what purpose was the dataset created?**

To analyze determinants of soccer players' market values, focusing on the Serie A. The goal is to enable robust insights for scouting, player valuation, and transfer market analysis.

2. **Who created the dataset and on behalf of which entity?**

The dataset was created by John Zhang as part of research into player market valuations.

3. **Who funded the creation of the dataset?**

The funding details are TBD.

4. **Any other comments?**

The dataset incorporates data from Transfermarkt, Stathead, FIFA, and Football-database.

2 Composition

1. **What do the instances that comprise the dataset represent?**

Each instance represents a Serie A player and their associated attributes, including market value, performance metrics, and contextual variables.

*Code and data are available at: https://github.com/Clearsky21z/Player_Market_Value_Analysis

2. **How many instances are there in total?**
The dataset contains 441 player records.
 3. **Is the dataset a sample or does it contain all possible instances?**
The dataset is a comprehensive sample of Serie A players for the 2023/24 season.
 4. **What data does each instance consist of?**
Each instance includes variables such as player name, age, position, market value, club and national team rankings, goals, assists, and minutes played.
 5. **Is there a label or target associated with each instance?**
The target variable is market value.
 6. **Is any information missing from individual instances?**
No, the dataset is complete.
 7. **Are relationships between individual instances made explicit?**
No explicit relationships are included.
 8. **Are there recommended data splits?**
No, but splits for training and validation can be derived.
 9. **Are there any errors, sources of noise, or redundancies?**
None identified.
 10. **Is the dataset self-contained?**
Yes, the dataset is self-contained and integrates data from multiple reliable sources.
 11. **Does the dataset contain confidential data?**
No confidential data is included.
 12. **Does the dataset contain offensive or sensitive content?**
No offensive or sensitive content is included.
 13. **Does the dataset identify any sub-populations?**
Sub-populations are identified by player positions.
 14. **Is it possible to identify individuals directly or indirectly?**
Yes, players are identified by name.
 15. **Does the dataset contain sensitive data?**
No sensitive data is included.
 16. **Any other comments?**
None.
-

3 Collection Process

1. **How was the data acquired?**
The data was scraped and collected from Transfermarkt, Stathead, FIFA, and Football-database.
 2. **What mechanisms or procedures were used to collect the data?**
A combination of web scraping and manual curation was used, scripts are available in the linked repository.
 3. **What was the sampling strategy?**
Players were sampled based on their inclusion in the Serie A for the 2023/24 season.
 4. **Who was involved in the data collection process?**
Employees for Transfermarkt, Stathead, FIFA, and Footballdatabase.
 5. **Over what timeframe was the data collected?**
Data was collected during and after the 2023/24 Serie A season.
 6. **Were ethical review processes conducted?**
TBD.
 7. **Did you collect the data from the individuals directly or from third parties?**
Data was collected from third-party platforms.
 8. **Were the individuals notified about the data collection?**
No notification was necessary as all data is publicly available.
 9. **Did the individuals consent to data collection?**
Consent is not required for publicly available data.
 10. **Was consent revocable?**
Not applicable.
 11. **Was a data protection impact analysis conducted?**
TBD.
 12. **Any other comments?**
None.
-

4 Preprocessing/Cleaning/Labeling

1. **Was preprocessing done?**

Yes, the data underwent rigorous cleaning to ensure consistency and reliability.

2. **Was the raw data saved?**

Yes, raw data was retained for future reference.

3. **Is the preprocessing software available?**

Yes, preprocessing scripts are available in the linked repository.

4. **Any other comments?**

None.

5 Uses

1. **Has the dataset been used for any tasks already?**

Yes, it was used in regression modeling to analyze market value determinants.

2. **Is there a repository linking papers or systems using this dataset?**

[GitHub Repository](#)

3. **What (other) tasks could the dataset be used for?**

Player scouting, valuation modeling, and market analysis.

4. **Does the dataset composition impact future uses?**

No significant limitations identified.

5. **Are there tasks for which the dataset should not be used?**

The dataset should not be used for non-sports-related analyses.

6. **Any other comments?**

None.

6 Distribution

1. **Will the dataset be distributed to third parties?**
Yes, under an open-access license.
 2. **How will the dataset be distributed?**
Via GitHub and linked repositories.
 3. **When will the dataset be distributed?**
It is already available.
 4. **Will the dataset be distributed under a license?**
Yes, an open-access license.
 5. **Have third parties imposed restrictions?**
No.
 6. **Do export controls or other regulations apply?**
No.
 7. **Any other comments?**
None.
-

7 Maintenance

1. **Who will support/host/maintain the dataset?**
John Zhang and his potential research team.
2. **How can the owner/curator/manager of the dataset be contacted?**
Contact details available in the linked GitHub repository.
3. **Is there an erratum?**
Not at this time.
4. **Will the dataset be updated?**
Future updates will include subsequent seasons.
5. **Are there retention limits?**
None.
6. **Will older versions of the dataset be supported?**
Yes, older versions will remain accessible.
7. **Is there a mechanism for extensions or contributions?**
Contributions are welcome via GitHub.

8. **Any other comments?**
None.

7.0.1 References

Please refer to the paper and repository linked above for further information.