Football Player Transfer Web Scraping

Lakshminarayanan (Laksh) Suryanarayanan

Shubham Kumar

Anirudh (Ani) Menon

Table of Contents

Executive Summary:

Football is a highly competitive sport, and the transfer market is an integral part of it. The process of acquiring players has become increasingly complex, and teams are now more focused on scouting and monitoring players' values and contract details from an early stage to gain a competitive edge. Our consultancy service aims to provide football clubs with the necessary resources to make informed decisions on player acquisitions.

We offer football clubs data on players and their corresponding transfer values from various leagues worldwide. We understand that keeping track of clubs and their respective players' contract details, market valuation, and other information can be time-consuming and resource intensive. Therefore, we provide real-time data on these players and their appropriate values, freeing up valuable resources for football teams.

Our approach is based on the transfermarkt web page, which is home to all the transfer details and data points that we collect to assign negotiation agents and scouts to monitor the players from the database we provide. Our service leads to more effective scouting and minimal delay in the allocation of travel for scouts to monitor particular players and negotiate with the team that these players play for.

Our consultancy services provide football teams with the necessary data and support to make informed decisions in the transfer market. Our objective is to help football clubs stay ahead of the competition and acquire the best players to improve their team's performance.

Background: Domain Knowledge and Industry Knowledge

Spectator sports have been a significant and lucrative business for decades. Fans worldwide demonstrate their commitment and passion for their favorite teams, making this industry one of the most followed and admired in the world. One of the most popular spectator sports in the world is football/soccer, which boasts an estimated fan following of approximately 4 billion people worldwide.

The popularity of football/soccer has led to a global fan base, with supporters from every nook and corner of the globe. As a result, football/soccer teams must sign the best players to compete and perform well on the field. The players' performances and attributes can determine a team's success and failure, and teams see these players as assets. A player's value can appreciate significantly with excellent performances, but it can also decrease with subpar performances.

The responsibility of acquiring these players falls on the director of football of each team. The recruitment process is an intricate process that involves scouting for potential players, valuing players based on their market worth, and negotiating contracts that align with the team's long-term goals.

While the football/soccer industry is highly competitive, the transfer market for players is still a manual process. Teams negotiate with each other to strike deals that benefit both parties. Negotiations can happen at any time during the transfer window, but teams tend to wait until the deadline expires to close a deal.

Business Problem

The process of player identification for transfers is a crucial component of the highly competitive world of spectator sports. It is a complex and intricate process that involves numerous parties, including players, clubs, and their agents. Despite advancements in technology, the process remains relatively manual, with communication between these parties being essential in striking a deal related to transfer fees, player positions, and contract details.

One of the major challenges in this process is the vast amount of information available online. While there are numerous websites that provide information about players and their transfer valuations, there is no streamlined procedure for collecting all the necessary details. As a result, teams are forced to spend an inordinate amount of time sifting through this information, which can be tedious and time-consuming.

Furthermore, the transfer market is highly competitive, with numerous clubs vying for the services of the most talented players. This has led to a bidding war, with clubs offering increasingly exorbitant transfer fees to acquire the services of these players. Given the high stakes involved, it is imperative that teams have access to all the necessary information about a player before making a transfer bid.

In addition to transfer fees and contract details, teams must also consider the player's position and style of play. Every team has a unique style of play, and it is essential that the acquired player fits into the team's system seamlessly. A misfit player can lead to a team's poor performance, which can be detrimental to their standing in the highly competitive world of spectator sports.

Furthermore, the value of players extends beyond their performance on the field. They are also valuable assets in the marketing activities of sponsors and companies. The fan following of a particular player can directly influence the way people perceive the products and services promoted by sponsors and companies. As such, it is crucial that teams take into account a player's marketability before making a transfer bid.

To address the challenges of the player identification process, there is a clear need to gather all relevant information into a centralized database. This database would provide teams with a comprehensive overview of a player's transfer valuation, contract details, playing style, and marketability. This would simplify the decision-making process for teams and ultimately lead to more successful and lucrative deals for all parties involved.

Data Source: Transfermarkt

The data source we are utilizing for our project is the renowned website, Transfermarkt. Since its inception in 2000, the website has been a one-stop shop for all football-related information, including player and team statistics, transfer fees, and player market values. It is widely regarded as a reliable source of information for various stakeholders such as fans, journalists, and analysts who are looking to stay abreast of the latest developments in the industry.

One of the key advantages of Transfermarkt is the comprehensive information it provides about football players. Our web scraping project focuses on gathering data about player performance, including goals and assists made, as well as relevant player attributes like age, nationality, height, and weight. Moreover, the website offers a valuation of players in the transfer market, taking into account various factors such as the player's contract status, age, performance, and other relevant metrics.

It is worth noting that the value of Transfermarkt goes beyond its utility for data gathering. The website has become an essential tool for many stakeholders in the football industry for a variety of reasons. For instance, businesses leverage the website to understand the market value of players they might be interested in signing, while journalists rely on it for accurate and timely reporting of transfer news. Moreover, fans also use the website to stay up-to-date with the latest player and team information, making it a valuable resource for personal interest.

The website's extensive information sources and data points demonstrate its significant and powerful influence on the football landscape. Its real-time updates and in-depth information have made it a valuable resource that continues to be appreciated by various stakeholders in the industry. Overall, our project aims to leverage this rich resource to help simplify and streamline the decision-making process of football teams during player transfers.

Explanation of the dataset

Understanding the URL structure is critical to accessing the relevant data for scraping and storing in our project. By knowing the different parameters within the URL, we can navigate to the specific player profile and extract the required data. Below are the details of each parameter:

1. Base URL: The base URL of the page is https://www.transfermarkt.com/.
2. Type of Page: The next part of the URL specifies the type of page being accessed, such as the team or player profile. As our project entails collecting player data, we will focus on accessing the player profile.
3. Player Name: The portion of the URL that contains the player's name is the "cristiano-ronaldo" part.
4. Profile: The "profile" part of the URL explains that this is the player's profile.
5. Spieler: The "spieler" part of the URL is the German word for the player, which is stored in the structure for historical reasons.
6. Player ID: The "8198" part of the URL is the particular number assigned to the player.

Web Scraping Routine

To scrape data from the Transfermarkt website and store it in MongoDB, several steps were taken. The following is a detailed description of each step of the web routine process.

1. Visited the URL:

The first step was to visit the Transfermarkt website's URL, which contained the information we needed. We accessed the URL: https://www.transfermarkt.us/spieler-statistik/wertvollstespieler/marktwertetop?ajax=yw1&altersklasse=alle&ausrichtung=alle&jahrgang=0&kontinent_id=0&land_id=0&page=1&spielerposition_id=alle, which displayed the top 500 valuable football players.

2. Goal:

Our goal was to extract the data of the 500 most valuable football players from the Transfermarkt website. To achieve this goal, we needed to scrape the data from all 20 pages of the search results. HTML data dump attached

3. Pagination:

The search results on the website were displayed with pagination, with 25 results per page and a total of 20 pages available for navigation. This meant that there were 20 pages of data to scrape. We utilized the pagination feature to extract the data from all 20 pages.

4. Looping:

To extract the data from all 20 pages, we used a loop to change the page number from the URL and fetch the data from all 20 pages. This was done to automatically navigate through all pages to retrieve the complete set of data.

5. Inspecting player details:

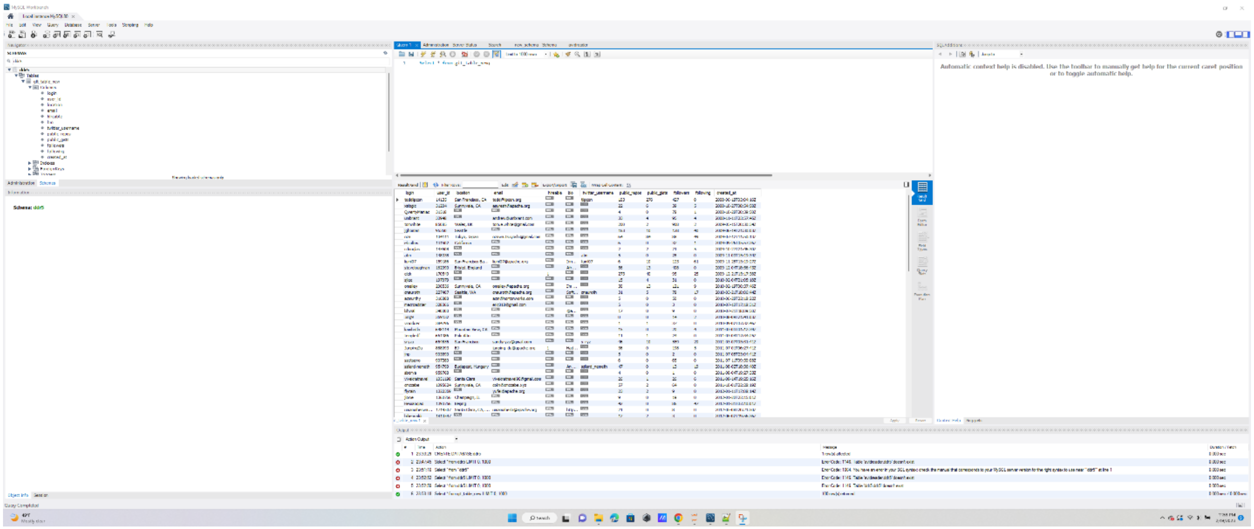
To extract the details of each player, we inspected each of the URLs and stored them as .htm files. These files were then attached to the zip file for future reference. The details we extracted for each player included their name, age, nationality, position, club, market value, and transfer fee.

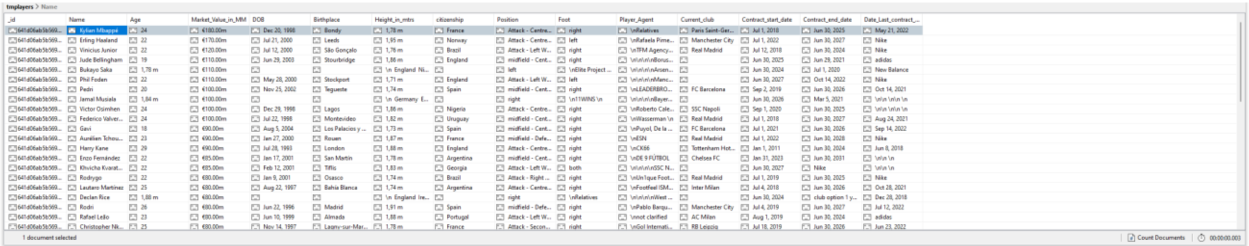
6. Creating a Python object:

From the files stored, we created a Python object containing the details of all the 500 players. This was done to make it easier to manipulate the data and prepare it for storage in MongoDB. The Python object contained the player details in a structured format, which allowed us to programmatically manipulate the data.

7. Storing data in MongoDB:

After creating the Python object, we stored the data in MongoDB. We first established a connection to the MongoDB client and then sent the data to a MongoDB collection. MongoDB was chosen for its flexibility and ability to handle large volumes of data.





tmplayers > Name

_id	Name	Age	Market_Value_in_MM	DOB	Birthplace	Height_in_mtrs	citizenship	Position	Foot	Player_Agent	Current_club	Contract_start_date	Contract_end_date	Date_Last_contract_...
641d06ab5b569...	Kylian Mbappé	24	€180.00m	Dec 20, 1998	Bondy	1,78 m	France	Attack - Centre...	right	\nRelatives	Paris Saint-Ger...	Jul 1, 2018	Jun 30, 2025	May 21, 2022
641d06ab5b569...	Erling Haaland	22	€170.00m	Jul 21, 2000	Leeds	1,95 m	Norway	Attack - Centre...	left	\nRafaela Pime...	Manchester City	Jul 1, 2022	Jun 30, 2027	Nike
641d06ab5b569...	Vinicius Junior	22	€120.00m	Jul 12, 2000	São Gonçalo	1,76 m	Brazil	Attack - Left W...	right	\nTFM Agency...	Real Madrid	Jul 12, 2018	Jun 30, 2024	Nike
641d06ab5b569...	Jude Bellingham	19	€110.00m	Jun 29, 2003	Stourbridge	1,86 m	England	midfield - Cent...	right	\n\n\n\nBorus...		Jun 30, 2025	Jun 29, 2021	adidas
641d06ab5b569...	Bukayo Saka	1,78 m	€110.00m			\n England Ni...		left	\nElite Project ...	\n\n\n\nArsen...		Jun 30, 2024	Jul 1, 2020	New Balance
641d06ab5b569...	Phil Foden	22	€110.00m	May 28, 2000	Stockport	1,71 m	England	Attack - Left W...	left	\n\n\n\nManc...		Jun 30, 2027	Oct 14, 2022	Nike
641d06ab5b569...	Pedri	20	€100.00m	Nov 25, 2002	Tegueste	1,74 m	Spain	midfield - Cent...	right	\nLEADERBRO...	FC Barcelona	Sep 2, 2019	Jun 30, 2026	Oct 14, 2021
641d06ab5b569...	Jamal Musiala	1,84 m	€100.00m			\n Germany E...		right	\n11WINS \n	\n\n\n\nBayer...		Jun 30, 2026	Mar 5, 2021	\n\n \n\n \n
641d06ab5b569...	Victor Osimhen	24	€100.00m	Dec 29, 1998	Lagos	1,86 m	Nigeria	Attack - Centre...	right	\nRoberto Cale...	SSC Napoli	Sep 1, 2020	Jun 30, 2025	\n\n \n\n \n
641d06ab5b569...	Federico Valver...	24	€100.00m	Jul 22, 1998	Montevideo	1,82 m	Uruguay	midfield - Cent...	right	\nWasserman \n	Real Madrid	Jul 1, 2018	Jun 30, 2027	Aug 24, 2021
641d06ab5b569...	Gavi	18	€90.00m	Aug 5, 2004	Los Palacios y ...	1,73 m	Spain	midfield - Cent...	right	\nPuyol, De la ...	FC Barcelona	Jul 1, 2021	Jun 30, 2026	Sep 14, 2022
641d06ab5b569...	Aurélien Tchou...	23	€90.00m	Jan 27, 2000	Rouen	1,87 m	France	midfield - Defe...	right	\nESN	Real Madrid	Jul 1, 2022	Jun 30, 2028	Nike
641d06ab5b569...	Harry Kane	29	€90.00m	Jul 28, 1993	London	1,88 m	England	Attack - Centre...	right	\nCK66	Tottenham Hot...	Jan 1, 2011	Jun 30, 2024	Jun 8, 2018
641d06ab5b569...	Enzo Fernández	22	€85.00m	Jan 17, 2001	San Martín	1,78 m	Argentina	midfield - Cent...	right	\nDE 9 FÚTBOL	Chelsea FC	Jan 31, 2023	Jun 30, 2031	\n\n \n
641d06ab5b569...	Khvicha Kvarat...	22	€85.00m	Feb 12, 2001	Tiflis	1,83 m	Georgia	Attack - Left W...	both	\n\n\n\nSSC N...		Jun 30, 2027	Nike	\n\n \n
641d06ab5b569...	Rodrygo	22	€80.00m	Jan 9, 2001	Osasco	1,74 m	Brazil	Attack - Right ...	right	\nUn1que Foot...	Real Madrid	Jul 1, 2019	Jun 30, 2025	Nike
641d06ab5b569...	Lautaro Martinez	25	€80.00m	Aug 22, 1997	Bahía Blanca	1,74 m	Argentina	Attack - Centre...	right	\nFootfeel ISM...	Inter Milan	Jul 4, 2018	Jun 30, 2026	Oct 28, 2021
641d06ab5b569...	Declan Rice	1,88 m	€80.00m			\n England Ire...		right	\nRelatives	\n\n\n\nWest ...		Jun 30, 2024	club option 1 y...	Dec 28, 2018
641d06ab5b569...	Rodri	26	€80.00m	Jun 22, 1996	Madrid	1,91 m	Spain	midfield - Defe...	right	\nPablo Barqu...	Manchester City	Jul 4, 2019	Jun 30, 2027	Jul 12, 2022
641d06ab5b569...	Rafael Leão	23	€80.00m	Jun 10, 1999	Almada	1,88 m	Portugal	Attack - Left W...	right	\nnot clarified	AC Milan	Aug 1, 2019	Jun 30, 2024	adidas
641d06ab5b569...	Christopher Nk...	25	€80.00m	Nov 14, 1997	Lagny-sur-Mar...	1,78 m	France	Attack - Secon...	right	\nGol Internati...	RB Leipzig	Jul 18, 2019	Jun 30, 2026	Jun 23, 2022

1 document selected

8. Data format:

Within MongoDB, the data was stored in .bson format. This format was chosen because it is efficient and optimized for the storage of large data sets. The .bson format allowed us to store the data in a compressed format, which reduced the size of the data and made it easier to query and manipulate.

These steps allowed us to successfully scrape data from the Transfermarkt website and store it in MongoDB for further analysis. The data can be used for various purposes, including identifying trends in player values, analyzing player performance, and developing predictive models for player transfers. By storing the data in MongoDB, we were able to easily manage and manipulate large volumes of data.

Database Design choices

The scraped data collected and stored in MongoDB offers advantages rather than using traditional relational database systems due to the following reasons:

1. Schema Flexibility: MongoDB allows for a flexible schema-free design to store and view player data which can be extremely useful if key data points are missing or not present in the transfermarkt website.
2. Data Scalability: Storing player transfer data in MongoDB allows for scaling horizontally, this improves performance and can handle larger amounts of data.

3. Speed: MongoDB is generally faster than traditional databases as it stores data in JSON format or document format which is optimized for data retrieval.
4. Complex Structures: As we continue to develop the scope of the project, the complexity of the structure offered by MongoDB can handle these complexities and be a boon for applications requiring hierarchical data models.

While relational databases prove to offer solutions where data is structured, the use of MongoDB can help store data efficiently and can expand the scope of the data scraped from the website.

Specifically, in the future, we want to add social media information as well to understand the media sentiment of each of the 500 players. Beyond a positive/negative sentiment, we wanted to assign a ‘Popularity’ score to each player. In addition to the attributes and contract details, the ‘popularity’ metric also influences a player’s potential moves based on the likeability among football fans. In our dataset, some of our players have multiple citizenships. MongoDB is best suited to capture it in a BSON or JSON format, as it returns all values for a single object. SQL wouldn’t be suitable as the fields would be dynamic, and as a result, wouldn’t fit in a tabular structure.

Business Efficiency of Created Dataset

The created dataset by web scraping player attributes and contract details would be used to answer various business-relevant questions in the field of football/soccer. The main answers that are provided through the dataset are as follows:

1. Player Valuation: The created dataset contains information on transfer fees and contract details such as the joining date and expiry date of the contract for each player. This subsequent data can be used to address questions related to the value of a particular

player and can help clubs make better decisions when it comes to the acquisition or sale of players.

2. Performance Analysis: The created dataset contains information about player attributes and their performance such as goals scored, assists provided, etc that allows teams to monitor the performance of players and analyze whether it is worth purchasing the respective player (MongoDB schema analysis in appendix)
3. Market Trends: The created dataset can be used to identify market trends such as a particular player's valuation and contract details. The teams can use this data to analyze and develop better strategies when it comes to transfers.
4. Scouting assistance: The created dataset can be used to assist team scouts by providing them with data regarding the players in the market and will allow scouts to target areas of travel to visit the players performing and understand their style of play. A bit of preliminary research is the idea here and can be expanded to allow the scouts to better understand the player and which results in efficient scouting.

Conclusion

In conclusion, this web-scraping project using Transfermarkt as the data source has opened doors for us to explore a wide range of possibilities in the football industry. With the valuable insights we have gained into player attributes and transfer details, we can address business problems and opportunities in the realm of football and spectator sports.

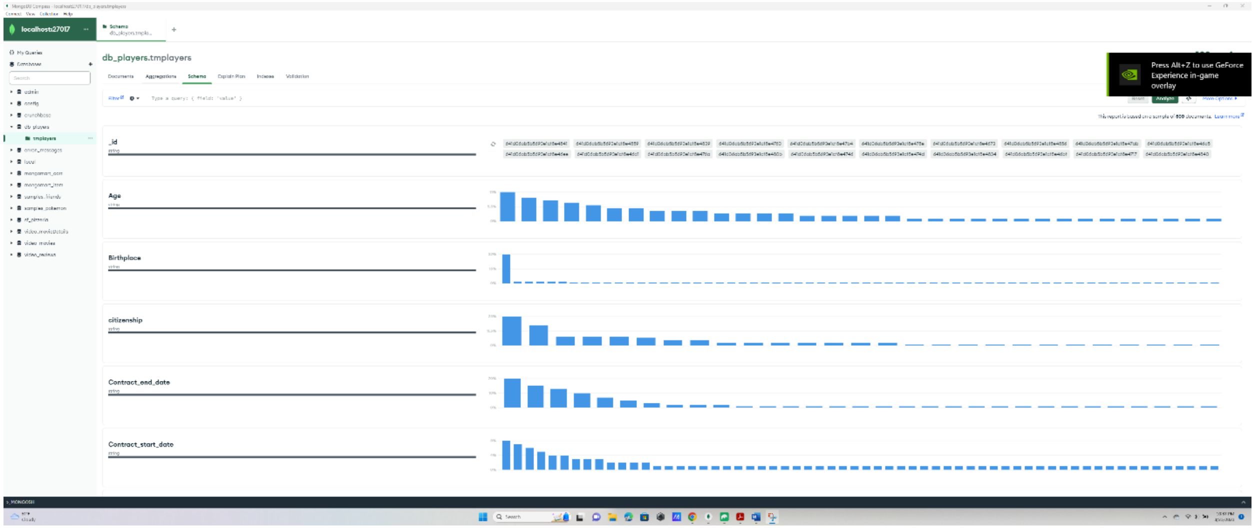
However, we must keep in mind the limitations of the scraping process and exercise caution while scraping data. The accuracy of the data available on the website and possible errors and inconsistencies must be taken into consideration while analyzing the data.

Moving forward, we can improve the accuracy and quality of data by implementing machine learning techniques and exploring and cleaning the dataset further. The potential sources of data and web-scraping ideas that can help reshape the recruitment and player understanding process in the football industry are immense.

Overall, this project has allowed us to gain skills in web-scraping and provided a wealth of information for football enthusiasts. We hope that the insights we have gained from this project can help us in making better decisions in the world of football and contribute to the growth and development of the industry.

Appendix

Analysis of the data from MongoDB



Press Alt+Z to use GeForce Experience in-game overlay

This report is based on a sample of 600 documents. Learn more