The Football Scout Bot

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering with specialty in Data Science

by

Riya Eliza Shaju

19BDS0061

Meghana Dirisala

19BDS0100

Muhammad Ali Najjar

19BDS0138

Under the guidance of

Dr. Gopalakrishnan T
SCOPE

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

March,2022

DECLARATION

I hereby declare that the thesis entitled "Football Scout Bot" submitted by me, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering with specialty in Data Science* to VIT is a record of bonafide work carried out by us under the supervision of Gopalakrishnan T.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date:26-04-22

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled "Football Scout Bot" submitted by **Riya Eliza Shaju (19BDS0061), Meghana Dirisala (19BDS0100), Muhammad Ali Najjar (19BDS0138)**, VIT University, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering with specialty in Data Science*, is a record of bonafide work carried out by him under my supervision during the period, 01. 01. 2022 to 30.04.202, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date :26-04-22 **Signature of the Guide**

Internal Examiner **External Examiner**

ACKNOWLEDGEMENTS

This is to acknowledge all those without whom this project would not have been reality. Firstly, I would wish to thank our Artificial Intelligence professor Mr. Gopalakrishnan T who gave his immense support, dedicated his time towards it and made us understand how to make this project. Without his guidance, the project would not have been complete.

While I was preparing this project file, various information that I found helped me in chapter Password security and I am glad that I was able to complete this project and understand many things. Through preparation of this Artificial Intelligence project was an immense learning experience and I inculcated many personal qualities during this process like responsibility, punctuality, confidence and others.

I would like to thank to my teachers who supported me all the time, cleared my doubts and to my parents who also played a big role in finalization of my project file. I am taking this opportunity to acknowledge their support and I wish that they keep supporting me like this in the future.

A project is a bridge between theoretical and practical learning and with this thinking I worked on the project and made it successful due to timely support and efforts of all who helped me.

Once again, I would like to thank my classmates and my friends also for their encouragement and help in designing and making my project creative. I am in debt of all these. Only because of them I was able to create my project and make it good and enjoyable experience.

Student Name

Executive Summary

Overview - The Quick Pitch

This project aims to create a product to reduce the boundaries and the efforts in finding out important information for transfers in football such as transfer values and player statistics to suggest realistic targets for teams based on their budgets.

Finding data from different sites to make analyses on whether a player would be affordable to a team and if they'd be a fit for the tactical system is a cumbersome process. We aim to reduce the steps involved in getting the information and implementing the process. There are no current non-professional tools available for the general public with interest in the sport that do what we do for free.

By implementing data-centric solutions on python using packages such as Selenium and Pandas to scrape from public databases so that we can create datasets to further analyze according to the needs of the application.

CONTENTS

.

List of figures

List of tables

Abbreviations

Sl no	Abbreviation	Full form
1	xG	Expected Goals
2	xA	Expected Assist
3	npxG	Non penalty expected goals
4	npxA	Non penalty expected assists
5	npxG+xA	Non penalty expected goals+assists
6	PK	Penalty kicks
7	CrdR	Red card count
8	CrdY	Yellow card count

INTRODUCTION

1.1. OBJECTIVE

Scouting is an important aspect of football. Scouts travel around the world to find players for a team that fit the team in terms of play style and budget. The idea of the project is to mirror the working of a real life scout by the use of datasets from different sites to get data on both players and teams and create a metric which matches players to a corresponding metric created to summarize the requirements of the entered team.

Create a code to parse through the players dataset and use selenium to further browse the selected player to get more information back from transfermarkt.com and then form the final dataset containing all of the required player information to move forward with the model.
Create a metric to summarize player statistics which will reflect all useful stats to in a comparable way so as to be able to cluster suitable players for a team to be interested in.
Finally, Create the bot and interface using Streamlit for easy interaction and increase readability and understanding of what is being shown in the final result

Our bot performs these tasks by studying datasets from websites that provide player performance statistics based on different metrics to create a metric to match to a similar corresponding metric which reflects the position and budgetary requirements of a team to further be studied and make further decisions to buy the player or not.

1.2 MOTIVATION

The average football fan can only rely on journalists and information released by clubs to find out if a player is suitable to a team's style and whether they are affordable. We ease these tasks by getting information using a custom bot to scrape websites for publicly available statistics.

The data is then combined and made available to check a player's availability based on contract information and whether they are suitable to a team's style of play. Tools like wyscout are only available professionally and not easily accessible for the general public for free.

1.3 BACKGROUND

Football is a simple sport. A match consists of 2 teams of 11 players each. The professional players might come at a youth level from local academies, academies in another city, country or continent, or on a professional level from another team. They are brought in by the team after being watched extensively by scouts for months. Scouts perform tasks like creating reports for coaches within the team to study and make further decisions.

Professional tools such as scoutpad and thescoutingapp exist but are not available for free. smarterscout is a tool which is available for free for limited access.

PROJECT DESCRIPTION AND GOALS

Our project is the result of decades of people wondering which players would be a fit in which teams' systems. Without a professional angle and unbiased articles from journalists, it is fairly difficult to come to a conclusion.

We hope to reduce the efforts in obtaining information for people to make their own final deductions using data obtained and processed based on player and team statistics.

The ‘Football Scout Bot’ is a project in its early stages and on further work and expansion can be made into a product with real financial value. The current goal is to suggest players for the entered team name by the user into the interface.

In the future, we can expand the project by involving metrics such as play style of managers and team and player histories to figure out close to accuracy if the player would be a proper fit into systems rather than them being a gamble by the club’s board.

Of course there are many aspects into this such as mental health and other personal and team dynamic issues that might hinder a player’s performances in a team which cannot be predicted, but we hope to find suitable analytic solutions to tackle these hurdles on our way to achieving our goal.

TECHNICAL SPECIFICATION

The project was built using the following python packages:

- numpy==1.21.5
- pandas==1.3.5
- requests==2.27.1
- streamlit==1.8.1
- vega_datasets==0.9.0
- openpyxl==3.0.9
- lxml==4.8.0
- Selenium==4.1.3

1. Selenium and requests were used to scrape the web sources to extract data which was further worked on using pandas and numpy to create useful datasets.
2. vega_datasets, streamlit, openpyxl and lxml were used to create the application interface and make it as visually informative as possible.

Post creation, the URL of the website is enough to run the application without having to download any softwares and can be used by anyone with the project link.

DESIGN APPROACH AND DETAILS

4.1 Design Approach / Materials & Methods .

A custom design approach made possible by scheduling out tasks and monitoring websites to figure out a way to successfully scrape the source and learn and use relevant python language packages to then work with the data to our requirement.

The metric created to analyze the player statistics is as follows:

Total Efficiency =Non-Penalty Efficiency - Red Cards - Yellow Cards

Non Penalty Efficiency

Expected Goals and Assists of the player subtracted by the Actual Goals and Assists of the player, giving a metric that suggests whether the player is efficient with his/her chances in front of goal and passes to put other through on goal or not

Expected Goals and Assists

What is xG?

Very simply, xG (or expected goals) is the probability that a shot will result in a goal based on the characteristics of that shot and the events leading up to it. Some of these characteristics/variables include:

Location of shooter: How far was it from the goal and at what angle on the pitch?

Body part: Was it a header or off the shooter's foot?

Type of pass: Was it from a through ball, cross, set piece, etc?

Type of attack: Was it from an established possession? Was it off a rebound? Did the defense have time to get in position? Did it follow a dribble?

Every shot is compared to thousands of shots with similar characteristics to determine the probability that this shot will result in a goal. That probability is the expected goal total. An xG of 0 is a certain miss, while an xG of 1 is a certain goal. An xG of .5 would indicate that if identical shots were attempted 10 times, 5 would be expected to result in a goal.

There are a number of xG models that use similar techniques and variables, which attempt to reach the same conclusion.

How xG is used

xG has many uses. Some examples are:

Comparing xG to actual goals scored can indicate a player's shooting ability or luck. A player who consistently scores more goals than their total xG probably has an above average shooting/finishing ability.

A team's xG difference (xG minus xG allowed) can indicate how a team should be performing. A negative goal difference but a positive xG difference might indicate a team has experienced poor luck or has below average finishing ability.

xG can be used to assess a team's abilities in various situations, such as open play, from a free kick, corner kick, etc. For example, a team that has allowed more goals from free kicks

than their xGA from free kicks is probably below average at defending these set pieces.
A team's xGA (xG allowed) can indicate a team's ability to prevent scoring chances. A team that limits their opponent's shots and more importantly, limits their ability to take high probability shots will have a lower xGA.

Penalty Kicks

Each penalty kick is worth .76 xG since all penalty kicks share the same characteristics. Comparing a player's goals from penalty kicks to their penalty kick xG can indicate a player's penalty kicking ability. Likewise, we can do the same for goalkeepers in these situations.

How xG total is calculated for a single offensive possession:

In some cases, a player or team's xG totals do not equal the sum of their shots. For instance, a team may attempt multiple shots in a single possession, but it is likely that these shots are contingent upon the outcome of the previous shot(s).

Take for example, a match between **Schalke 04 and Nürnberg**:

In the 78th minute, Nürnberg attempted three shots which ultimately led to a goal. Hanno Behrens attempts a shot that is saved, but he is able to take a second shot as the ball is deflected off the defender. The second shot goes off the woodwork, which allows Adam Zreľák to easily tap it in. According to StatsBomb's expected goals model:

Behrens' first shot with the goalkeeper in his way = .37 xG
Behrens' second shot with the goalkeeper out of position but a defender in the way = .68 xG
Zreľák's shot with an open net = .81 xG
The sum of these three shots is 1.86 expected goals, even though it is impossible to score more than one goal in a single move. To solve this problem, we find the probability that the defending team does not allow a goal in this possession. In this case, the calculation is:

(1 - .37) x (1 - .68) x (1 - .81) = .0383 or a 3.83% probability that Schalke does not allow a goal.

To find Nürnberg's xG, we simply subtract that probability from 1:

1 - .0383 = .9617 xG

In other words, we estimate that an average team in a similar situation would be expected to score a goal 96.17% of the time.

We use a similar method when calculating xG for individual players. Adam Zreľák receives .81 xG from his single shot while Hanno Behrens receives:

1 - (1 - .37) x (1 - .68) = .7984 xG

This shows why a team or player's total xG may not equal the sum of the xG from their shots and why a team's total xG may not equal the sum of the xG from their players.

Possessions that include a penalty kick

Similarly, we include shots taken from a rebound after a penalty kick with xG from penalty kicks. Take this Alexis Sanchez penalty kick for example:

As mentioned above, the penalty kick attempt = .76 xG
The second shot after the rebound, from 6 yards and with the goalkeeper unrecovered from the save = .72 xG
Since the second shot is a result of the first, we use the same probabilistic method in the previous example. Rather than a total 1.48 xG (.76 + .72), the calculation is:

1 - (1 - .76) * (1 - .72) = .9328 expected goals

However, since the second shot is also considered to be a part of the penalty kick xG, Sanchez gets 0 npxG (non-penalty expected goals) on this play.

Update: On July 31, 2020, StatsBomb upgraded their xG model with the inclusion of shot impact height, which is the height of the ball when a shot is struck. For many shots in which we know the height of the pass preceding the shot, there will be little to no impact. For the other shots, however, there is a "sizeable impact" on their xG value. This update improves the quality of an already industry-leading xG model.

What is xA?
xA, or expected assists, is the xG which follows a pass that assists a shot. This indicates a player's ability to set up scoring chances without having to rely on the actual result of the shot or the shooter's luck/ability. Note: Because xA comes from passes, not all assists will be given an xA value.

Red Cards and Yellow Cards

The red card is used by the officials to remove a player from the match. It means the automatic ejection of the player and that the player's team will remain shorthanded for the remainder of the match. Red cards can be given if the same player has received two yellow cards in the same match.

In essence, a yellow card is given as a caution or warning. It provides players receiving them another chance to stay on the field for the remainder of the game, whereas a red card means that the player has to leave the pitch with immediate effect.

These give an idea of the player’s on field disciplinary record.

4.2. Constraints, Alternatives and Tradeoffs

A. Current constraints in the project include a minimal dataset which can be made bigger by considering bigger datasets from sources which will in turn cost more time and power in scripts.
B. Some of the tradeoffs we have included in the project is that the suggested players are currently the same for every team but that can be fixed by using bigger datasets and additional code to accommodate them.

SCHEDULE, TASKS AND MILESTONES

1) Find out useful websites to scrape using scraping tools and custom python scripts.

2) Create a code to parse through the site and use selenium to further browse the selected player to get more information back from transfermarkt.com and then form the final dataset containing all of the required player information to move forward with the model.

3) Create a metric to summarize player statistics which will reflect all useful stats to in a comparable way so as to be able to cluster suitable players for a team to be interested in.

4) Finally, Create the bot and interface using Streamlit for easy interaction and increase readability and understanding of what is being shown in the final result

5) Our bot performs these tasks by studying datasets from websites that provide player performance statistics based on different metrics to create a metric to match to a similar corresponding metric which reflects the position and budgetary requirements of a team to further be studied and make further decisions to buy the player or not.

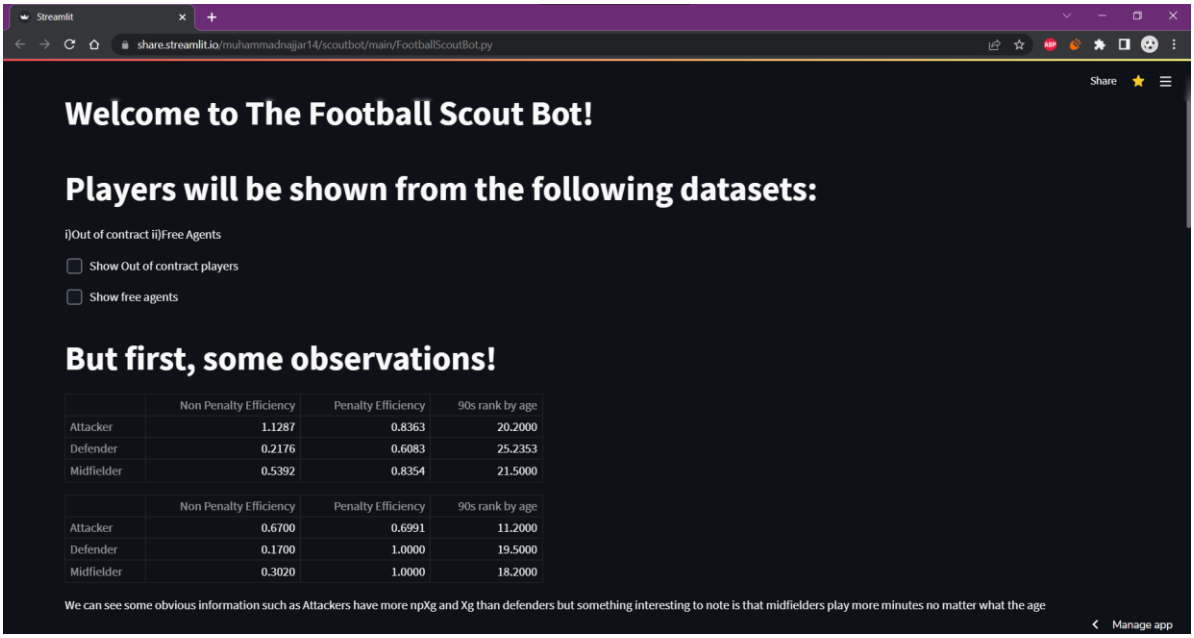
PROJECT DEMONSTRATION

URL TO PROJECT:
https://share.streamlit.io/muhammadnajjar14/scoutbot/main/FootballScoutBot.py

URL TO PROJECT GITHUB:
https://github.com/muhammadnajjar14/scoutbot

Screenshots of project UI :

1. Initial View:



2. On clicking ‘Show out of contract players’:

3. On clicking ‘Show free agents’:

Players Available For Free:

Players Available For Free

	Player Name	Position	Previous Team	League	Age	Height	Preferred foot	Free since	Value	Primary Natior
0	Rémy Cabella	Attacking Midfield	FK Krasnodar	Premier Liga	32	1,71 m	right	Mar 9, 2022	€8.00m	France
1	Edgar Ié	Centre-Back	Trabzonspor	Süper Lig	27	1,82 m	right	Jan 14, 2022	€7.50m	Portugal
2	Robert Beric	Centre-Forward	Chicago Fire FC	Major League Soccer	30	1,88 m	right	Jan 1, 2022	€2.50m	Slovenia
3	Kwang-song Han	Centre-Forward	Al-Duhail SC	Qatar Stars League	23	1,78 m	both	Jul 1, 2021	€1.70m	Korea, North
4	Charlie Daniels	Left-Back	Colchester United	League Two	35	1,78 m	left	Jan 20, 2022	€1.60m	England
5	Mateo Musacchio	Centre-Back	SS Lazio	Serie A	31	1,82 m	right	Jul 1, 2021	€1.50m	Argentina
6	Jürgen Damm	Right Winger	Atlanta United FC	Major League Soccer	29	1,85 m	right	Feb 25, 2022	€1.50m	Mexico
7	Jota Peleteiro	Right Winger	Deportivo Alavés	LaLiga	30	1,78 m	left	Jul 1, 2021	€1.50m	Spain
8	Mohamed Diamé	Defensive Midfield	Al-Ahli SC	Qatar Stars League	34	1,84 m	right	Jul 1, 2021	€1.20m	Senegal
9	Afriyie Acquah	Central Midfield	Al-Batin FC	Saudi Professional League	30	1,79 m	right	Jan 30, 2022	€1.20m	Ghana

4. Observations made and interactive visualizers created on analyzing data:

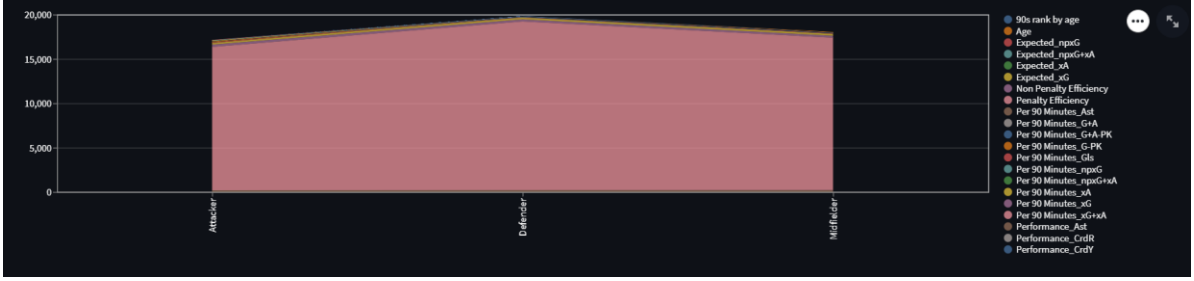
But first, some observations!

	Non Penalty Efficiency	Penalty Efficiency	90s rank by age
Attacker	1.1287	0.8363	20.2000
Defender	0.2176	0.6083	25.2353
Midfielder	0.5392	0.8354	21.5000

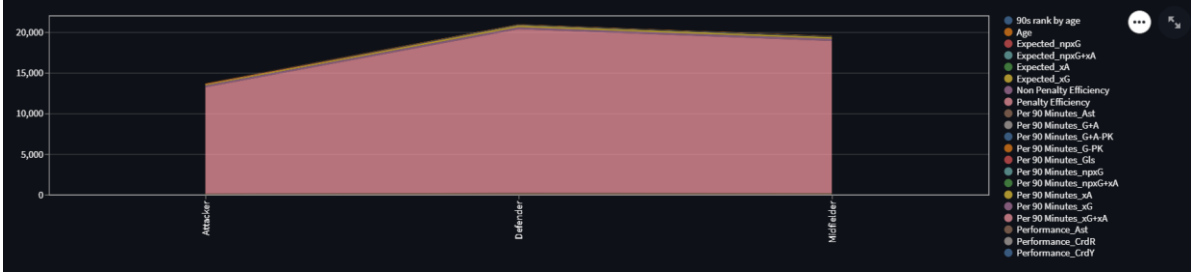
	Non Penalty Efficiency	Penalty Efficiency	90s rank by age
Attacker	0.6700	0.6991	11.2000
Defender	0.1700	1.0000	19.5000
Midfielder	0.3020	1.0000	18.2000

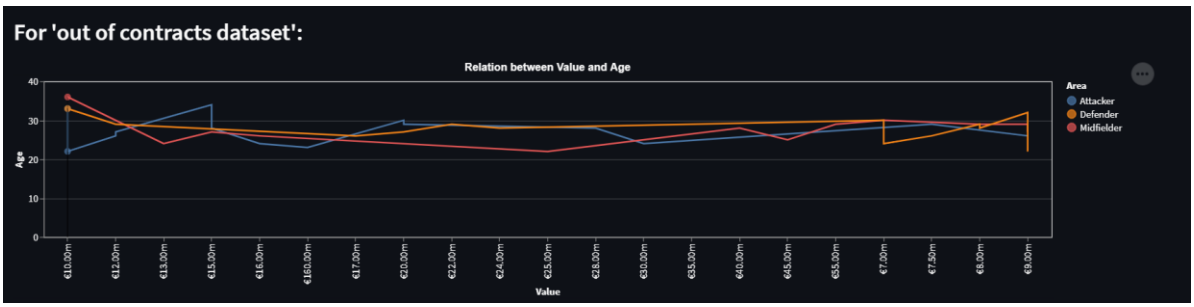
We can see some obvious information such as Attackers have more npXg and Xg than defenders but something interesting to note is that midfielders play more minutes no matter what the age

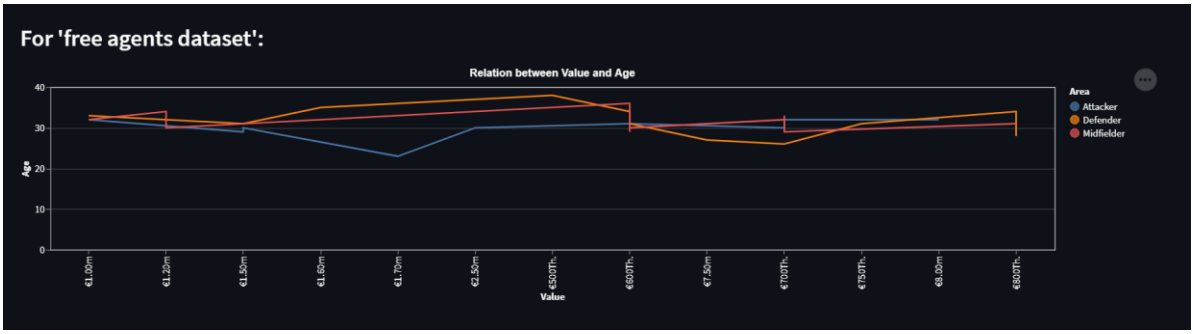
For 'out of contracts dataset':



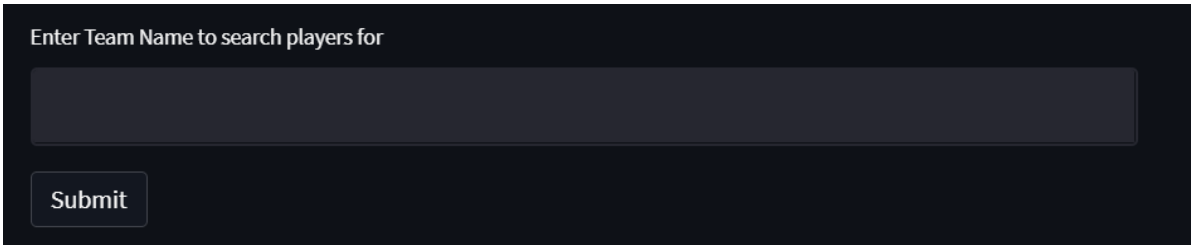
For 'free agents dataset':







5. Search bar to search for team and display budgetary information:



with example:

6. Showing suggested players with relevant information :

For out of contract players:

Suggested players (players not already in picked team):

Show Out of contract players

Best Players Going Out Of Contract Names, Details and Statistics:

	Player Name	Position	Current Team	League
33	Dan-Axel Zagadou	Centre-Back	Bor. Dortmund	Bundesliga
7	Boubacar Kamara	Defensive Midfield	Marseille	Ligue 1
28	Eddie Nketiah	Centre-Forward	Arsenal	Premier League
0	Kylian Mbappé	Centre-Forward	Paris SG	Ligue 1
43	Amos Pieper	Centre-Back	Arm. Bielefeld	Bundesliga
5	Ousmane Dembélé	Right Winger	Barcelona	LaLiga
15	David Brooks	Right Winger	Bournemouth	Championship
20	Sean Longstaff	Central Midfield	Newcastle	Premier League
2	Franck Kessié	Central Midfield	AC Milan	Serie A
31	Enis Bardhi	Attacking Midfield	Levante	LaLiga

For free agents:

Best Players Available For Free:

	Player Name	Position	Previous Team	League
3	Kwang-song Han	Centre-Forward	Al-Duhail SC	Qatar Stars Le
20	Marc Navarro	Right-Back	Watford FC	Premier Leag
1	Edgar Ié	Centre-Back	Trabzonspor	Süper Lig
14	Prince Gouano	Centre-Back	Amiens SC	Ligue 2
15	Donny Toia	Left-Back	Real Salt Lake City	Major League
24	Ramon Azeez	Central Midfield	Granada CF	LaLiga
29	Jordy Delem	Defensive Midfield	Seattle Sounders FC	Major League
6	Jürgen Damm	Right Winger	Atlanta United FC	Major League
18	Ahmed Khalil	Centre-Forward	FC Shabab Al-Ahli Dubai	UAE Pro Leag
30	Bryan Pelé	Left Midfield	AEL Limassol	Protathlima

RESULT

The result of the project's functions is an accurate dataset describing suggested players for the team name entered. This can be further expanded as mentioned before by including more metrics to be considered in the final decision process such as manager, team and player history which can open up an entire new dynamic for dealing with how players are considered for purchase by a club's board and ownership.

REFERENCES

1. https://smarterscout.com/about
2. https://www.thescoutingapp.com/
3. https://aiscout.io/
4. https://metrica-sports.com/?gclid=CjwKCAjwsJ6TBhAIEiwAfl4TWH841iFBCdR6DB4mK03ZKMUJ9wLN2i4ShSYcML7urUA1ByDKvBaPjBoCkVUQAvD_BwE
5. Fbref.com/en/players
6. https://www.transfermarkt.com/statistik/vertragslosespieler
7. https://www.transfermarkt.com/statistik/endendevertraege
8. https://adblockplus.org/
9. https://dataminer.io/
10. https://streamlit.io/