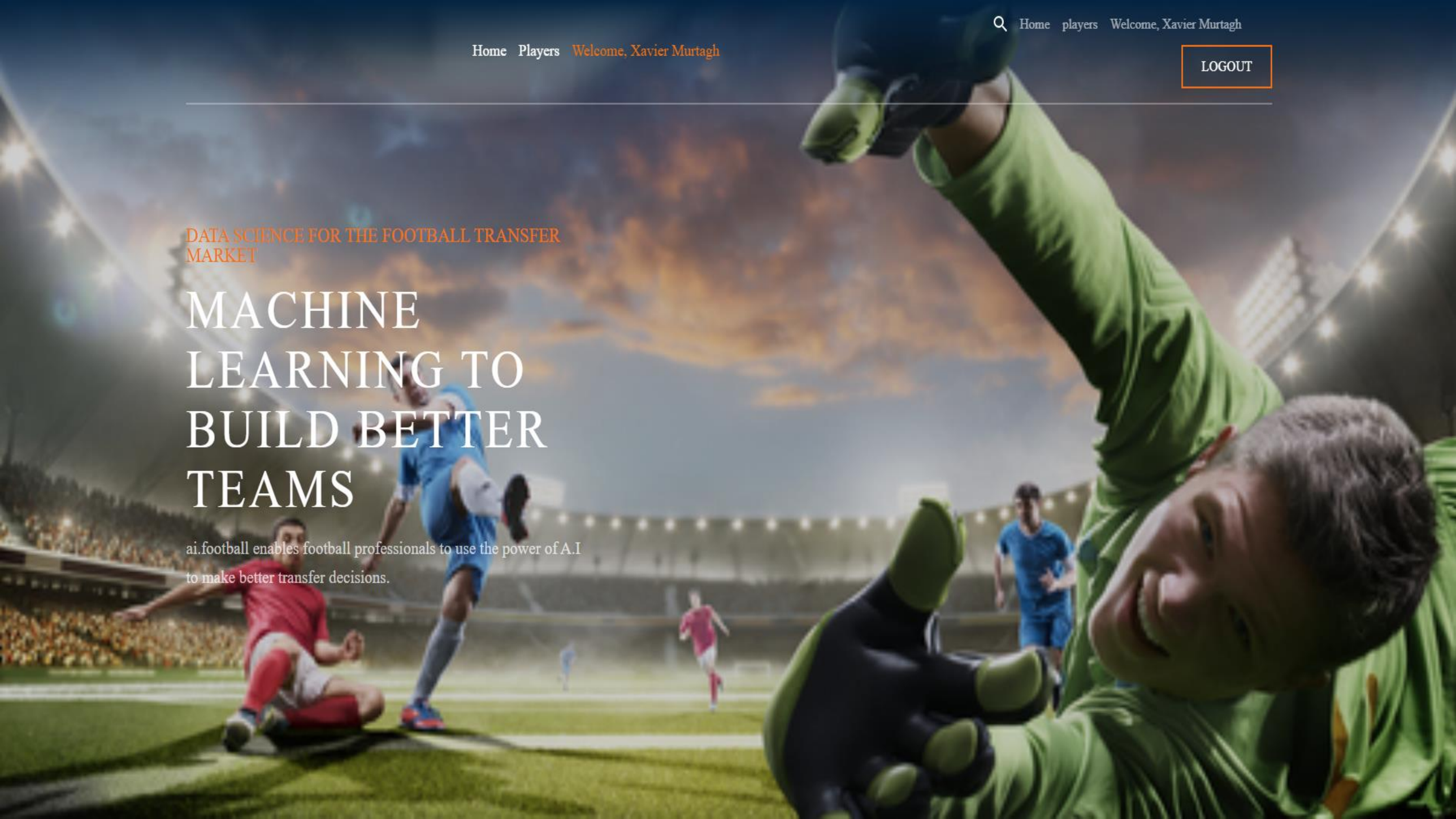

Home players Welcome, Xavier Murtagh
LOGOUT
Home Players Welcome, Xavier Murtagh
DATA SCIENCE FOR THE FOOTBALL TRANSFER MARKET
MACHINE LEARNING TO BUILD BETTER TEAMS
ai.football enables football professionals to use the power of A.I to make better transfer decisions.

TABLE OF CONTENTS

INTRODUCTION

ANALYSIS

DOCUMENTATION & DESIGN

DEVELOPMENT

TESTING

EVAULATION

APPENDIX

ANALYSIS

PROBLEM IDENTIFICATION & RESEARCH

ANALYSIS – INTRODUCTION

Sports teams are multi-million and in some cases multi-billion pound businesses. Professional teams need to continually succeed. They need to win to keep sponsors happy or to get more money from them, to get more money from television rights(the higher you finish the more money you get) and to raise ticket sales. But all sports teams are only as good as the players in their team.

As a keen football fan I keenly followed the headlines surrounding the summer football transfer market. The news headlines were full of stories about players being transferred for millions. These stories were then followed by other articles about the financial problems that many of the largest football teams in the world are currently facing. As part of my research I learned that football teams in the Premier League lost £1 billion in the 2020/21 season mainly thanks to the ongoing COVID-19 situation but they still set a record amount in expenditure last year which was over £1bn. Another example is the financial problems that Barcelona FC currently face. They have a debt of £1.2 billion and this summer, they had to sell Messi one of the greatest football players ever and sell a number of players for much less than they would normally have been sold for in order to help balance their finances.

I began to think - is there a better way that football clubs could use artificial intelligence and in particular machine learning to help build better teams more cost effectively?

The aim of my program:

‘To develop a machine learning system that enables football clubs to better assess if a player is over or under valued relative to his or her peers. It is not fantasy league, but a real world problem that could potentially save clubs millions of pounds every year”.

ANALYSIS – RESEARCH, METHODS & SOURCES

In comparison to some other sports such as baseball, football has been quite late with starting to collect large sets of data for analytics purposes due to the difficulty of recording this data with past technology. Despite the vast amount of money in football, I believe that football is still run very traditionally and it has only been in the last few years that new approaches to training, nutrition and developments have been developed in football. This also applies to the hiring and scouting of new players. It has been the traditional approach to rely on football scouts (often older men) who go to watch players at football matches and those people with experience in professional football. However, my approach is best summed up by the story of the successful football manager who never played professional football in his career who responded to criticism over his lack of experience with his famous quote when he said: “I never realised that to be a jockey you had to be a horse first.”

There is also a growing football analytics industry that is beginning to use a wide variety of AI techniques, especially: computer vision, statistical learning and game theory. Players need to take decision-making when surrounded by other players and as such game theory (which is the theory of interactive decision making), becomes important. Players can be tracked and game scenarios can be recognised automatically from widely-available image and video inputs. This information can be used by AI tools to make better decisions.

The rapid progress in artificial intelligence (AI) and machine learning has opened many possibilities. Deepmind is a relatively new UK based company, but as part of my research into AI and machine learning I watched the documentary AlphaGo. It showed how their computer scientists were able to use machine learning to build a system that was able to defeat the world’s best Go player, something that until recently was previously was thought to be almost impossible and 10 years before experts predicted that AI could complete such an achievement.

ANALYSIS – RESEARCH, METHODS & SOURCES

Problems Faced by User

In the analysis for my project, I researched several books, websites and articles. I recently read an article in the Financial Times about Billy Beane who is part of a consortium that is buying Barnsley FC. This article showed me that my project could help any team no matter their size. I had also watched the film Moneyball move as part of my research and knew he was the pioneer of using sports statistics in sports.
Below is a referenced copy from the article:

Moneyball guru eyes beautiful game with Barnsley takeover.

“US baseball’s Billy Beane hopes to transfer his statistical strategy to football. Baseball guru Billy Beane is part of a consortium that is close to acquiring a 98.5 per cent stake in the Yorkshire football club. Billy Beane, the US baseball executive behind the statistically savvy approach to sport known as “ Moneyball”, is looking to test his ideas on the beautiful game in God’s own country. Billy Beane's ideas are credited with helping to transform Major League Baseball in the US The US executive is one of the pioneers behind “sabermetrics” in baseball, using statistical insights to gain hidden advantages in the game, such as identifying undervalued players to acquire them cheaply. As immortalised in Michael Lewis’s bestselling book, Moneyball, the Oakland A’s went on a record 20-game winning streak in 2002 under Mr Beane’s leadership, despite fielding a team with one of the lowest wage bills in Major League Baseball.“ There are fundamental differences between baseball and football,” said Ben Marlow, an executive at 21st Club, a football consultancy that has advised the new owners of Everton and Swansea City on recent takeovers. “But the principle that you can use information that is evidence-based, instead of more traditional subjective methods, to gain an advantage is something that is transferable between sports,” he said. **“Football has been slower to adopt some of these ideas that have been adopted in other industries.**” – Financial Times

ANALYSIS

STAKEHOLDERS, USERS & MODELLING OF THE PROBLEM

Premier League clubs splash out £1.1billion during summer transfer window

Arsenal are the biggest spenders (close to €150 million on five players)!

Juventus Didn't Just Lose Ronaldo. It Took a $17 Million Hit Too

All right, it is harsh to call them the worst. But, there is no doubt that Manchester City overpaid for the talented Grealish. His market value would be no more than €70 million. This deal is especially surprising since City did not lack depth in his role.

Big-spending Arsenal may have made the biggest blunder of all.

Loser: Arsenal

If the objective of this summer's transfers was to the infuse the squad with young talent, Arsenal overpaid for players. If the objective was to bring in top-tier talent, Arsenal failed to do so. Ben White is a talented young defender, but is he worth £50m when Varane and Dayot Upamecano went for around €40m and Kurt Zouma went for £25m? Aaron Ramsdale is a backup goalkeeper the Gunners paid £24m for. **Martin Odegaard is a good signing** based on last season's performance but is a slight overpay at €35-40m. Albert Sambi Lokonga and Takehiro Tomiyasu come in for €20m each while in-house youngsters like William Saliba, Reiss Nelson and Ainsley Maitland-Niles go on loan. Arsenal also surprisingly completely avoided a solid free transfer pool. Spending money doesn't guarantee results, especially when it's spent on second-tier talent.

ANALYSIS – USERS & STAKEHOLDERS

From a small selection of news headlines you can clearly see that football clubs are spending millions on players but not always wisely. Take for example Eden Hazard, he played seven successful seasons scoring 110 goals in 352 games for Chelsea. He won two Premier League titles, two Europa Leagues, the FA Cup, and League Cup. In the summer of 2019, he signed a 150 million contract for Real Madrid. This transfer was one of the most expensive transfers involving an English club in the history of football (https://www.telegraph.co.uk/football/2019/06/07/eden-hazard-leaves-chelsea-real-madrid-move-worth-130m/). However, the transfer cannot be considered successful. The market value of Eden Hazard has decreased significantly since 2019. In June 2019, it was 150 million, while, in December 2021, it was only 20 million. Why do some football clubs overpay for players, why do some teams overperform despite not having the same financial resources and is there a way to see predict if players are going to keep their value better?.

Users & Stakeholders

The primary stakeholders and users of the system are football clubs, but sports teams are now run as businesses so a user could be the club's Chief Financial Officer who is responsible for managing club finances. Teams like Brentford FC or Barnsley FC (where Billy Beane has been brought in to advise) have smaller budgets and therefore data science gives them any statistical advantage over their opponents that they can try to exploit.

Problems Faced by Users & Stakeholders

A single football match can generate an enormous amount of data: goals, attempts, attempts on goal, corners, yellow and red cards, time on field of each player, substitutes, ball possession, and many others. However, is every piece of information valuable? Or, on the other hand, are there any other types of data that can be useful in terms of player evaluation, transfer and team building?

ANALYSIS – MODELLING OF THE PROBLEM

The Problem

The problem can be modelled in one sentence; is there a computational solution that can help football clubs in evaluating the accurate value of a football player?

This is very difficult to answer in one simple formula. There are numerous factors that influence the price such as the physical characteristics of the player, the financial situation of the club buying or selling the player and even how many shirts of the player that the club can sell. However, a good model can determine a player's value, based on their skills. Each on-pitch position has an influence on the statistics that a player can be measured by. A good goalkeeper should have more team clean sheets and a forward should have more goals. The league a player plays in is also an important factor. You would expect that a Man United player should be worth more than a Hampton FC footballer.

The Solution

My solution will source data on football players. This data will contain metrics on the player, such as goals scored, but also other metrics including age. I then propose to also source data on the currently estimated value of a player. By combining how well a player performs and comparing to their value I want to build a model to determine if a player is over or undervalued using their performance metrics.

Machine Learning – Over or Under Valued

The idea of my system is similar to what happens in the financial markets. Analysts gather lots of metrics about a company and enter this into a machine learning algorithm. The algorithm then produces an estimate if the share price is over or undervalued and then buys or sells. My system works in a similar way. The data won't be in real time, but I propose to build a database of players going back over three seasons. Then using this 'training' data use a neural network algorithm to determine if the value of a player is over or under valued.

ANALYSIS – OBJECTIVES OF THE SOLUTION

Structure of the Project

The project is divided into three sections. The first part is the data analytics which is designed to provide data science analytics that football clubs can use to get better player insights. The data is taken from the dataset that includes players from top European leagues. The analytics provide the manager with valuable insights into distributions, importance of age, player's position etc.

The second part of the data analytics focuses on the machine learning process and determining the value of the players. Models are created depending on the players' position, with each containing different variables.

The third part of the project is the website. This uses the data to create a website that the user can log into to view the details of players. The website user will then be able to see the values of the players and compare to those calculated by the machine learning algorithm

Objectives of The Project

Further in this Analysis section I will define specific objectives and measurable success criteria by which the project can be measured by. However, I would like to already define the high-level objective of this project and that is to develop a system that has a real-world application.

Player transfers are one of the most important components in building a successful sports team. By taking the 'Moneyball' data science approach of using data to help clubs identify over and under valued players. It uses the machine learning techniques, and it has been tested on the real-life data. But this project goes further than using simple statistics. It uses machine learning and the power of neural networks to help identify a system that helps football clubs make better transfer decisions which is something that could potentially save them millions every year.

Target User

The users of the system will be primarily football club managers, but also anyone who is involved in football club transfer market. This can include finance managers who are responsible for the finances of the club.

ANALYSIS – USER INTERVIEW

User Interview

I was able to contact the football manager of a smaller football club to interview him regarding my proposed system and to find out if this system would be useful in him and his club . The following is a summary of the responses he gave to my questions.

Club Manger

Q. Do you currently use data science in any aspect of running the club. If so please describe further.

“We are a small club and don’t have the resources for a team of data science analysts like the large clubs. However, football is a game full of statistics and we are beginning to explore this more.”

Q. How do you manage the process of football transfers?

“Our club has a scouting division that manages the initial process of identifying players that may be suited to our club. This focuses on both the junior players, but also more experienced players. The scouts present their analysis to me as the manager. We then discuss with the owners those players that we have identified as potential targets.

Q. Do you use a third party agent to manage the process?

“Football agents are becoming almost as powerful as the players! We have to deal with agents particularly when dealing with more experienced players. As our budget is smaller than the bigger clubs we try to bring as many players as possible through our academy system.”

Q. How do you determine the value of a player to buy?

“This is very difficult. There are so many variables that influence a player’s value. The obvious factor is their skills, but there are so many other factors. Their age, their experience, the leagues that they have played all matter. There are also external factors which are very difficult to factor. COVID is an example. A lot of clubs suffered during this period and had to sell some players just to balance the books.”

ANALYSIS – USER INTERVIEW

Q. How do you ascertain the value of a player you wish to sell?

“Again, very difficult. A player may get approached by a larger club and we will have no choice but to sell. Smaller clubs always have to sell their best players to survive. The value is determined by the buying club, though I often wonder how they actually work out the price they are willing to offer. It always seems somewhat random to me. Sometimes they offer a higher price than the market and sometimes under.”

Q. Do you use third party websites or data analytical providers?

“I know of a few websites. They can be useful as they have some useful, but it is limited as it is just a big list really. There is no way compare one player versus another. Something like that would be really useful. We could use this in our meetings,

Q. Would a system that compared player values and used Artificial Intelligence to ascertain if a player was over or undervalued be of use?

“I profess I don’t know anything about artificial intelligence, but if there was such a system this would be a very helpful. We don’t have much apart from our experience in doing this for years. If there was a system that we could use that would help our decision making on whether a player’s transfer market value was over valued, that would be great.

Q. Are there other analytical tools that would be potentially useful?

“Yes, there are many. There are so many stats for every player that it can be overwhelming. If there was a way to visualise the statistics that would help me a lot. I think in pictures more than numbers so anything that helped paint the picture would certainly help”

Q. How do your scouting team present their player reports?

“They write these on Word. They include some basic tables, but nothing too complicated.”

Q If there was a player report that you combined all the data analysis that I mentioned would be this helpful?

“Very. That sounds like a very useful report. If it showed information about the player in a better format than Word then that would be a big improvement. It would be useful to see how they compared to another player. If there was such a system then I would certainly use it. It could maybe save our club some money”

ANALYSIS – RESEARCH & EXISTING SOLUTIONS

Data science in the sports market is undergoing massive growth. During my research I found that that there are analytics websites for all types of sports including baseball, basketball, cricket, rugby and football. Games such as football manager have interestingly also been used for scouting as well as fantasy football websites. In my research I have identified some existing solutions, but also identified their limitations.

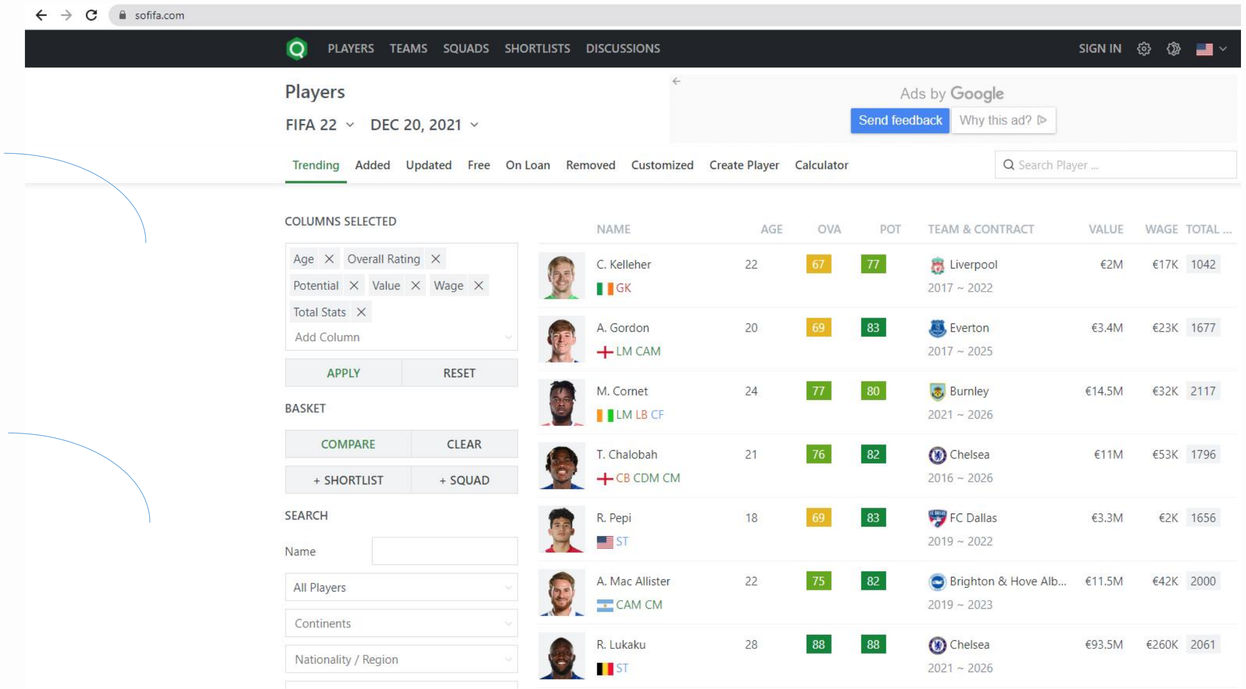
.

Limitations of Existing Solution

The sofifa.com website has useful player data but does not provide any 'value added' analysis. I summarise it more like a large, but simple database. Football clubs would not find that this information very insightful as it simply lists basic analytics. There is no ability to assess if a player is over or under valued. I have also noted below the style of the sofifa.com website. It is not user friendly and is very cluttered. My first impression of this website is that it looks dated.

This is the front page which is very unwelcoming. It is very crowded. I prefer the google approach which is a simple front page that clearly states what the site does.

Website only has details of player transfers, but no comparative analysis to show if this player is over or under valued



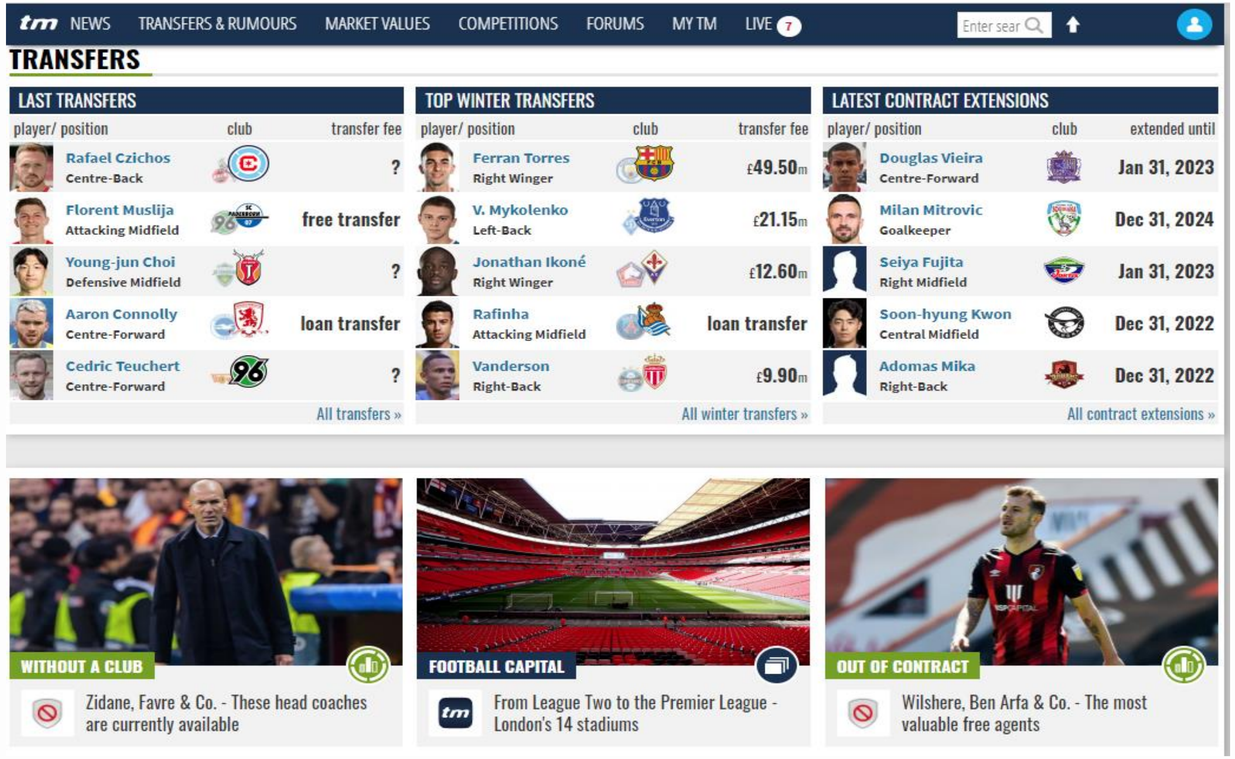
ANALYSIS – RESEARCH & EXISTING SOLUTIONS

The other existing solution that I found during my research is tranfermarkt.co.uk. This is fanbase website that is free to use. It is based in Germany and also has information on the football transfer market.

Limitations of Existing Solution

The limitation of this site is that it again does not give the football club the ability to determine if the transfer value represents good value. The transfer fee is listed, but there is nothing to explain a players market value. However, this website will be one of the primary data sources in building the data analytics for this project. The challenge which is outlined in the Design section will be to build the code to extract the data from this site, cleanse it into and useable format and then build the data analysis!

Cluttered home page



Website only has details of player transfers, but no comparative analysis to show if this player is over or under valued

ANALYSIS – RESEARCH & EXISTING SOLUTIONS

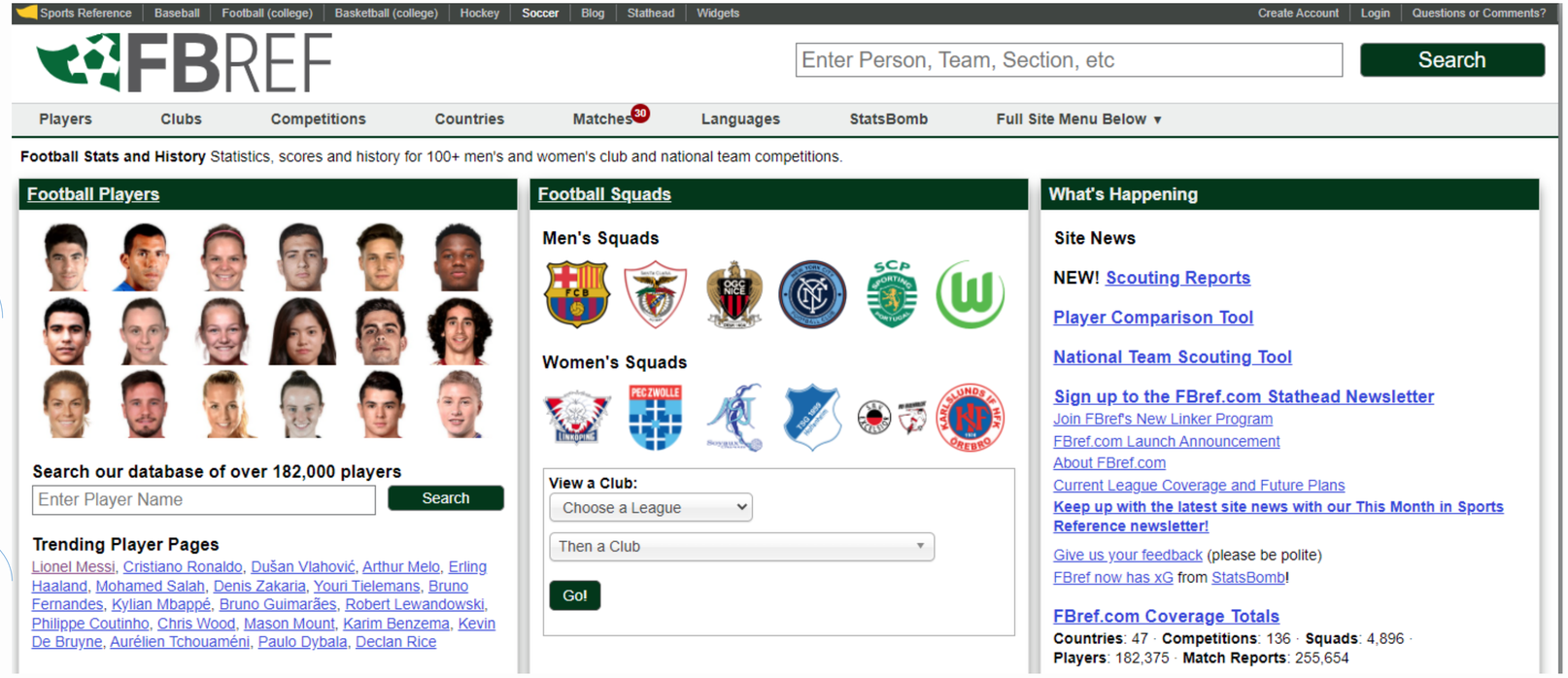
Another existing solution that I have found in my research is the site fbref.com. It is the most comprehensive of all the existing solutions for detailed player analytics.

Limitations of Existing Solution

The key limitation of this site is that it does not contain transfer market data. Also, like all the other existing solutions there is no 'relative value analytics' and therefore little use to a football club who is spending tens (and sometimes hundreds) of millions on choosing one player over another.

Another cluttered home page

Dated layout



Just lots of old style html links

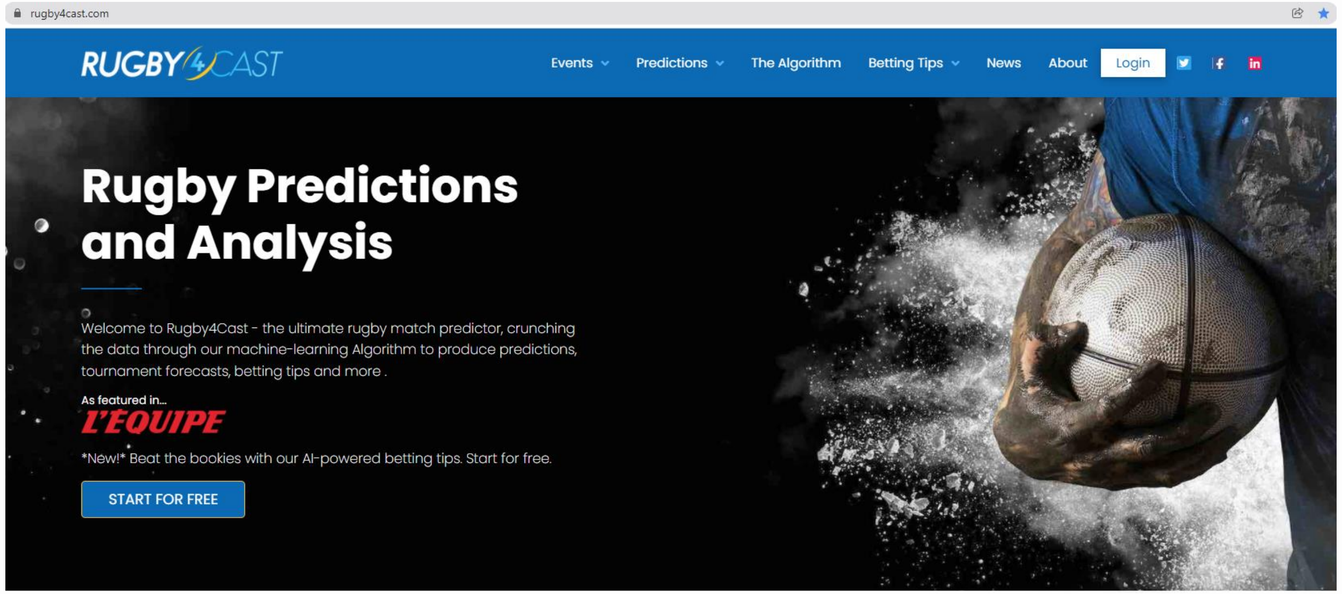
ANALYSIS – RESEARCH & EXISITING SOLUTIONS

Football isn't the only sport where you can see an increasing reliance of computational analytics. My research uncovered the site www.rugby4cast.com. It says on its homepage that it is "the ultimate rugby match predictor crunching data through our machine-learning algorithm".

Summary - how my system differs from fantasy league or sports betting sites

This site introduces an important point to note at this stage in the analysis which is to note what my system is not designed to be.

1. My system isn't about trying to predict results of a game. There are an increasing number of betting sites that use AI (though I am not aware of any that have worked very well).
2. My system is designed to be a real world solution using in-depth statistical analysis system for comparing relative values based on real data My system is not designed to be a fantasy league. It is real data collated from the actual transfer market and then analysed. The football club will be able to use it to make real decisions about real players in a real business that could help save them a lot of money.



ANALYSIS – OVERVIEW OF SOLUTION REQUIREMENTS

Assessment of Coding Language

In order to develop a complex system, I need to use a high-level programming language. The language needs to be able to handle complex data manipulation and produce graphical and analytical output. There are several languages which focus on functions rather than objects, in these programming languages each decision follows on from the previous decision. These are useful for simpler linear outcomes. However, complex interfaces are now built using object orientated programming languages.

AI and Machine Learning are the areas which have some of the most interesting developments in computer science. Many of these projects are powered by python. There are many advantages to using python. It offers an object orientated structure, but it is flexible to present it as procedural code. The python language is efficient, reliable, and much faster than most modern languages. Python can be used in nearly any kind of environment, and you will not face any kind of performance loss.

One further advantage of the python language is that it can be used in many varieties of environments such as mobile applications. If I had more time, I would use python to let me roll out the system across multiple platforms.

Assessment of Solution Requirements for A.I

Another reason why I chose python as the core development language was because I will be using python's machine learning capabilities. The design and coding sections will go into more detail about the python algorithms.

Assessment of Coding Language for Database & Website

The development of the website will be in html and the back end will be MySQL. I will used phpMyAdmin to manage the database for the interaction with the website. These are opensource development languages that are used in most database driven websites.

ANALYSIS – HIGH LEVEL OBJECTIVES & SUCCESS CRITERIA

In this analysis section of the project, I want to list the high-level objectives of the project and to define specific objectives to assess in the Evaluation section if I have achieved these . The target users of my system are football club managers, but the system could be adapted to any sports team. of objectives follows. These objectives will be tested, and the output will be assessed to see if my system meets the objectives.

Section 1 – Data Analytics & Machine Learning

1. Data Acquisition & Cleansing

Data is the single most important variable in any system. Even if you had a perfect model without good data the system would have little value. The first challenge that I have identified is that there is no simple downloadable data set containing all the data on player and transfer market values. My objective will be to source data from multiple sources, cleanse it and then have it in a format that will be able to be used for the machine learning and the website.

2. Data Science & Analytics

The csv file of rows of players and columns of attributes is in of itself little use. Yes, it is the basis on which the system will be built, but for the user the requirement will be to see comparative data and graphical analytics in a professional report.

3. Machine Learning

The objective will be to use a neural network to see if the transfer valuation of the player represents good value. The objective will be to build a comprehensive data set, then apply linear regression to the key factors that influence a player's value. The neural network will produce a precited value of the player. This will enable the club to make better decisions when buying and selling players.

Section 2 – Website & Database

The system will provide the users with advanced data analytics, but another objective will to be develop a modern website. The design of the website will be professional to appeal to the target user. The current websites that I identified in my research are dated, not user friendly and have limited use as they don't provide advanced data analysis or machine learning.

ANALYSIS - OBJECTIVES & SUCCESS CRITERIA

1. Data Acquisition & Cleansing

Objective	Process	Success Criteria
Sourcing website that has market values for football players. The data has to be free to use and cannot be obtained in breach of copyright.	Obtain transfer market website to obtain player transfer market values. Thorough analysis of html, data structure, URLs and output required. No simple 'download all data', therefore web scraping script will have to be developed	Csv output file of player ID's and their respective market value.
Sourcing websites that have football player statistical attributes. The data has to be free to use and cannot be obtained in breach of copyright.	Obtain fbref's player attributes. Thorough analysis of html, data structure, URLs and output required. No simple 'download all data', therefore web scraping script will have to be developed	Csv file of player ID's and their respective statistical attributes.
Create database of players across multiple leagues.	In order to build a system that accurately analyses several data points, it is necessary to get player data across the top 5 European football leagues.	Data output file has data not just from one league but from the following 5 leagues: (i) English Premier League (ii) Spanish La Liga (iii) German Bundesliga (iv) Italian Serie A (v) French Ligue 1
Create dataset of player attributes across multiple seasons.	In order to build a system that accurately analyses several data points, it is necessary to get player data across multiple years and seasons.	Data output file has data not just from one season, but from the following seasons: (i) 2018/19 (ii) 2019/20 (iii) 2020/21
Develop a comprehensive data set of all the player attributes, transfer values for multiple seasons across multiple leagues.	As there are several csv output files each sourced from different sites. The data needs to be consolidated and player attribute field names need to be matched.	Success of the data acquisition and cleansing process will be defined by the output. This output will be a consolidated csv of: (i) Player attributes (ii) Player values (iii) Over multiple leagues (iv) Over multiple years

ANALYSIS - OBJECTIVES & SUCCESS CRITERIA

Data Science & Analytics

Objective	Process	Success Criteria
Produce a graphical output to show the most valuable players	Develop code to create a chart of valuable players	Easy to read chart which will be used in the Data Analytics report
Produce a graphical output to show the histogram of players and frequency of values	Create a histogram and log charts	Easy to read chart which will be used in the Data Analytics report
Player comparison – make a report that graphically compares two players	Use the data collated to show the merits of one player vs another.	Graphical pie chart and table

Machine Learning

Objective	Process	Success Criteria
Develop a linear regression model	Using the linear regression apply to the different positions: Goalkeepers Defenders Midfielders Forwards	Each of the player positions will have a separate output and their performance assessed relative to their peer group
Build a neural network to assess if the output from the linear regression shows that a assess if a player's value is over or under valued.	Assess suitable neural network. Train with test and training data set. Apply weights for prior seasons Produce a predicted value	The predicted value of the player will give the user the ability to know if a player's transfer value is higher or lower than it should be given the performance of that player in key factors.

ANALYSIS - OBJECTIVES & SUCCESS CRITERIA

Website & Database

Objective	Process	Success Criteria
Domain	Find a good web domain that represents what I want to achieve with the system. It should ideally highlight the link between the analytics and football	The website URL needs to be modern and for the user to know what system is going to offerabout
Hosting	Setup hosting with domain hosting company	Contract with hosting company to host website
Professional Design	Develop wireframe to set out the website template. Design key components that highlight to the user what the service can provide	Create a user friendly website that delivers a professional service and acts as a marketing tool for the system
Website Schema	Develop in html & MySQL the following pages: 1. Front Page 2. Login Page 3. Player List Page 4. Player A.I Page	The user will be able to load the website on the web and be able to navigate it easily
Home page	Design professional home page	Simple but effective home page that states clearly what the site does
Database	Develop database of players and users	
User registration and login	Codify user login and registration pages	User will be able to register and login
Search for a particular player	Use database search function to find player and display details on the page	The success will be that a user's search will show the player details
Player page with A.I value	Once the user has selected the player the page will load. The A.I estimated transfer value will be given and an estimate of how the player is over or under valued relative to their current transfer value	The machine learning process will calculate a value depending on how well has performed and then estimate if the transfer value currently listed for the player is too high.

ANALYSIS – LIMITATIONS OF PROJECT

In this section I already want to highlight some of the limitations of the project. There is a no perfect system. Even Microsoft is continuously fixing its software. In the Testing phase I will expand on these further, but it is important to introduce some of the issues that I have encountered in the initial analysis phase.

1. Frequency of Data

As will be discussed in more detail in the design section, data is the most important aspect to any data science and machine learning process. The phrase ‘data is the most valuable resource in the world’ is said for good reason. Google and many other global companies are in their simplest form only data mining. Their values are linked to the quality of their data, and it’s the same with my project. The primary limitation of the system is the age of the data set. There are so many up-to-date statistics that I have not been able to capture that would be useful in creating a better model. For example, real time analysis on a player’s performance. The dataset I am using is generated from the 20/21 season, but to make the system better, data would need to be updated after every game. It would take a lot of data scientists to capture this information and a lot of computing power to analyse it. Machine learning is only as good as the continual updated dataset it can learn from. A good example of this data limitation would be a player’s current injury status and how prone they are to injury as this would significantly impact the value of a player.

2. Data Set

The dataset will have thousands of records over several leagues but excludes international competitions. International games such as the world cup can have a significant impact on the value of a player. For example, Jack Grealish was one of the best players for England at the recent Euro Championships. His transfer value to Man City was more expensive as a result of these performances. The system is therefore limited by not including all the statistics of international games.

ANALYSIS – LIMITATIONS OF PROJECT

3. Accuracy of Data

The data set I am using is free and from several public websites, but another limitation of the system is that I am not able to determine if the data is accurate. There is no independent audit, and the collection of the data may be prone to error. One incorrect variable can have significant impact on the results.

4. Outliers in the Data Set

Another limitation is that within the dataset there are many data outliers. For example, there are several players that do not play often and therefore there is a lack of a complete dataset for the player throughout the season. These outliers have an impact on the analysis. To help address this issue I have had to cleanse the data. This is an accepted approach in data science otherwise outliers have an erroneous impact on the results. This is discussed in further detail in the design section.

5. Model

In addition to the previously mentioned issues on the quality of the data, so is there one with the actual model. My idea for the project was brought to life by all the articles and headlines on the losses incurred by football clubs as a result of them over-paying for players. It is a real-world problem and one that involves billions. Even if the dataset was perfect, to then build a model that would model everything perfectly would be unrealistic. However, as I will show in the following design and testing sections, the model I have developed has a very good base to build on.

6. Machine Learning

Machine learning algorithms are more and more used to solve problems that only a few years ago would not have been possible. Self-driving cars is an example. I have designed the system to use a neural network to take the data input and produce an output that predicts if the player is over or undervalued. This will provide the football club user with a useful metric. However, there are limitations as to what this can do. The output from the neural network may or may not be correct. The user should therefore only use this as an additional factor to consider.

ANALYSIS – LIMITATIONS OF PROJECT

8. Database

The database has limitations as it only has a fixed dataset. It is not continually updated with live data. Another important limitation of the database is that although MySQL is widely used it does have limitations. There are many new database technologies that can manage a lot of data quicker.

7. Website

The design of the website will be to develop a site that is professional. The dataset that will be used is comprehensive, but the amount of information that can be displayed on the website will still be limited. The design section outlines the key pages that I want to develop, but there will be limitation as to all the pages that I am able to develop in this version. There will be other pages in future versions which will have more details and analysis.

DESIGN

DATA ACQUISTION & PROCESSING

DESIGN – DEVELOPMENT SYSTEMS

Anaconda

Anaconda is a distribution of the python programming languages and is frequently used for scientific computing (data science, machine learning applications, predictive analytics, etc.), as it simplifies package management. Within the Anaconda environment I was able to use JupyerLab

JupyterLab

JupyterLab is a web-based environment for notebooks, code, and data which helps visualise your outputs nicely. The Jupyter Notebook App is a server-client application. This can be executed locally or can be installed on a remote server and accessed through the internet. Its flexible interface allows me to develop code and analytics for the data science and the machine learning elements.

Visual Studio

I have also used Visual Studio as a programming environment for the python, PHP and HTML code.

Neural Networks

Sikit-learn is the neural network that I used for the machine learning modelling. This will be described in further detail later in this section.

Database

MySQL will be the database. I have chosen this as this is opensource and free to use. MySQL is used in many of the database driven websites and therefore is a reliable system for development.

Website & Domain

The website will be hosted on a leading hosting web server. The domain will be purchased and will use HTML with phpMyAdmin to administer the MySQL database I will develop a modern website.

Swimlane Diagram

The design section is structured according to each of the projects four components:

(i) Data Acquisition

(ii) Analytics

(iii) Machine Learning

(iv) Website

For each of these sections there will be a data flow charts, algorithms and supporting Pseudo-code

In the analysis section I identified the three core components of the project, the data acquisition, the data analytics and then the machine learning. These are then all used as the basis for the website which enables the user to access not only the database of player analytics, but a unique A.I tool to make the unique assessment as to whether a football player is over or under valued in the transfer market.

A swim lane diagram is a logic flow of the processes involved in the data science. Each of the three lanes illustrate the user, the computation and the output.

DESIGN – DATA ANALYSIS

Introduction

The most important part of any machine learning system is the data. Unfortunately, there aren't any publicly available easily downloadable databases with all the information I need for the data analytics. The challenges are therefore as follows:

1. Research all possible data sources and assess suitability
2. Develop code that would acquire the data
3. Consolidate all the data sources into one csv
4. Cleanse the data to ensure it is suitable for data analysis
5. Perform statistical analysis

Data Sources

As highlighted in the Analysis section, my in-depth research identified two main sources:

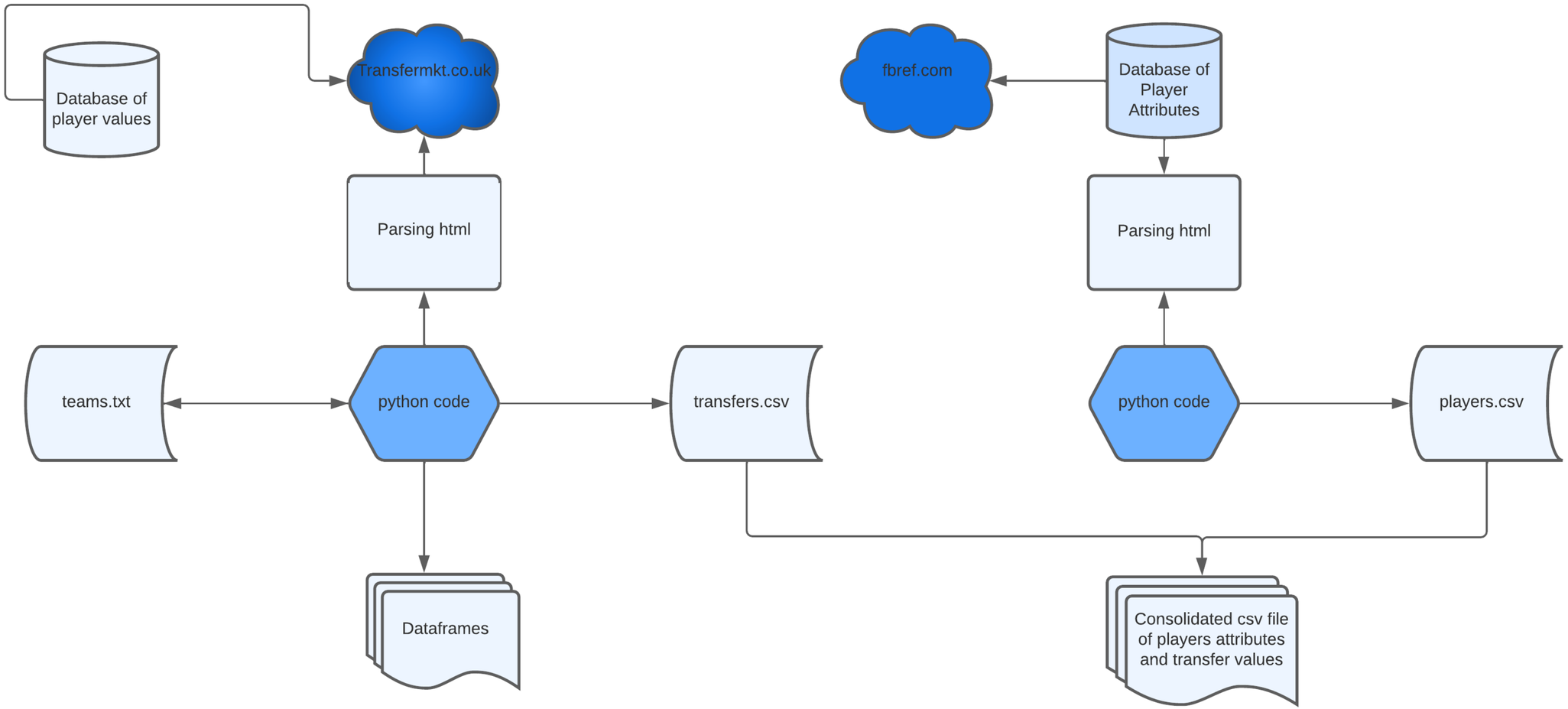
i. fbref.com - which contains lots of statistical information on football players. I can download the data from this website as a csv file and this will contain the statistics of the players that will be used for the data analytics. However, this website does not contain any information on the transfer value of a player and therefore I need to find additional transfer data to build my model.

ii. Transfermarkt.co.uk – this site does contain the player transfer values, but there are two problems. The first is that it does not contain many player statistics that I need, and the second issue is that I am not able to download any of the data from the website. Therefore, I need to develop code that will scrape the data from the website in a process called data scraping.

It is important to note that the websites used are free to use and are not paywall protected. Clear reference is made throughout this project to the sources of the data. See Appendix.

DESIGN – DATA ACQUISTION PROCESS FLOW

The following diagram summaries the process to acquire the data and output in consolidated csv scraping process



DESIGN - DATA ACQUISTION & PROCESSING

Develop code to acquire data “Data scraping”

As I am not able to easily download the transfer value of a player from the website, I need to develop the code to do so. I will then use this data and combine it with the player data I already have, to create a ‘consolidated.csv’ file. I then need to repeat this process for each season.

Data Cleaning Algorithm

The master dataset is very large, so I have created a subset which will be most applicable to the users I identified in the Analysis section. I have only selected the top professional leagues and to avoid ‘outliers’ that may impact the results, I have decided to remove all players that aren’t in this range:

FOR all_players
SELECT league{La Liga—Spain, Serie A—Italy, Bundesliga— Germany, English Premiership}
Exclude player_value <1,000,000
AND games_played <5

Data Analytics

The result of the data scraping process is a csv file of thousands of records. I will then begin the next part of the system which is to generate data analytics that will enable the user to achieve the ‘”Goal to build better teams”.

DESIGN – DATA SCRAPING ALGORITHMS

Parsing HTML

HTML is made up of elements, like a paragraph or a link, that tell the browser what to render. For data scraping, I will use this information to tell my code what information is needed. The next stage is to inspect the page structure. Transfermarkt.co.uk is a German website and has an unusual mix of German and English. In order to parse HTML, I need to use python for parsing HTML. It creates a parse tree for parsed pages that can be used to extract data from HTML.

Next, I have created a variable called ‘headers’ and assign it a string that will tell the website that we are a browser, and not a scraping tool. This is very important otherwise the browser will not let me do it. Next, the first one assigns the address that we want to scrape to a variable called ‘page’. When I have the page, I need to define the page headers. Once I have the headers, I can extract the content. The final stage is to parse the website code into html. I am then be able to search through this for the data that I want to extract. Then I will save this page output.

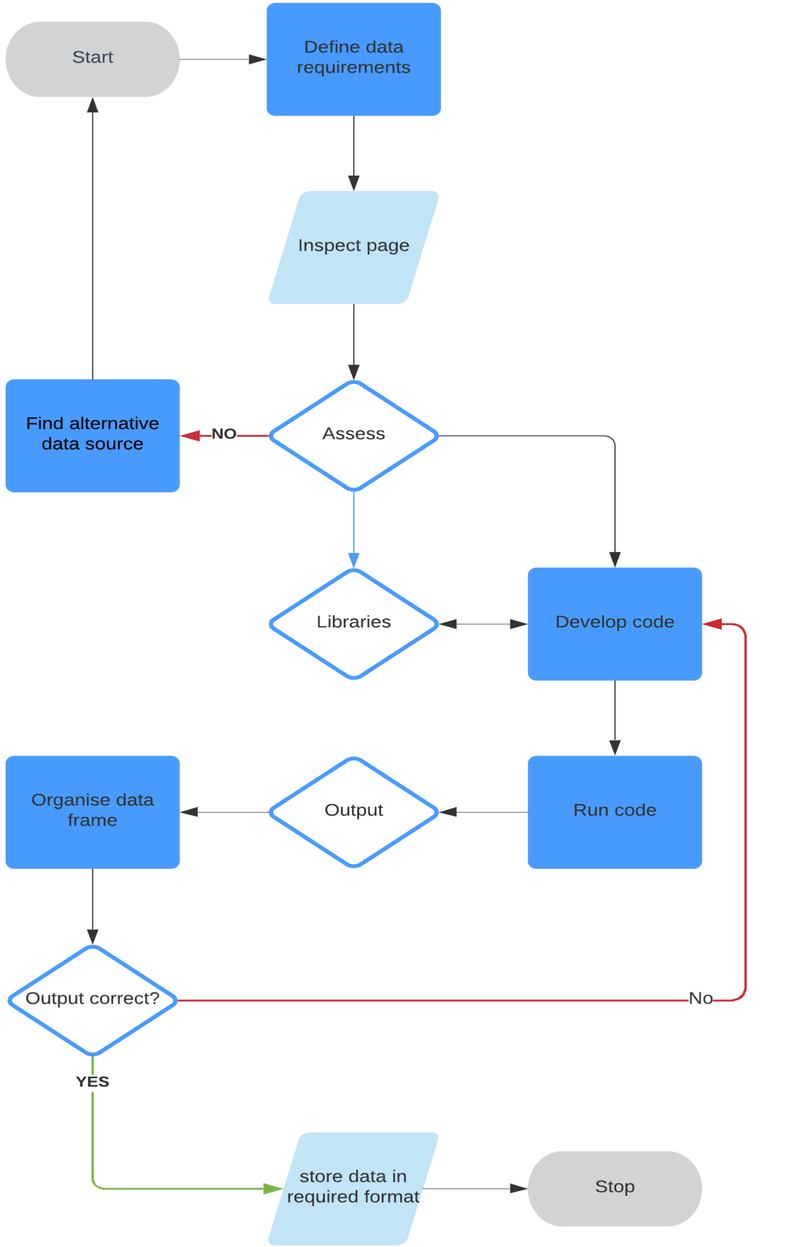
```
Parsing HTML
Import html page parser
#headers routine required so that the website knows we are a real user and not a bot
        headers = {'User-Agent‘'}
        page = {URL of website}
        pageheader = get (page.headers)
        pagecontents = get (page.content, 'html.parser’)
        pageoutput
```

DESIGN – DATA SCRAPING ALGORITHM

Introduction

When I run the code for data scraping, a request is sent to the URL. In response, the server sends the data and allows me to read the HTML page. The code then parses the HTML page, finds the data and extracts it.

1. Confirm URL with required data
2. Inspect the page to find the data needed to be extracted
3. Develop the code
4. Reference library
5. Run the code and extract the data
6. Organise the data into a data frame
7. Store the data in the required format
8. Repeat the process



DESIGN – DATA SCRAPING CLASS DEFINITIONS & DATA FRAMES

Class Definitions - Data Scraping

A python class is used to create a new object. Every time a class object is instantiated, a new object is initiated. The class objects that I create can be used again whenever needed. The player's name is a link. This is denoted as an ‘a’ tag in HTML, so I will use the ‘find_all’ function to look for all the ‘a’ tags in the page. I can use the class given to the players’ names specifically on this page to only take these. The class name is passed to the ‘find_all’ function as a dictionary. This function will return a list with all elements that match my criteria.

Players = find_all

I need to find the class name for the players on the website

"a", {"class": “players"})

I want to output the first name in the Players list which is:

Players[0].text

Output: 'Jack Grealish’

However, the transfer values are not a link, so I need to develop a subroutine to identify them by. They are in a table cell, denoted by ‘td’ in HTML.. I now need to assign this to Values.

Values = find_all("td", {"class": “values "})

Values[0].text

The output is: '£105.75m’

As there are 25 players in the list, so I need to use a for loop to add the first 25 players and value to new lists and create a new data frame with them:

PlayersList = [] ValuesList = [] for i in range(0,25):

PlayersList.append(Players[i].text) ValuesList.append(Values[i].text)

df = pd.DataFrame({"Players":PlayersList,"Values":ValuesList}) df.head()

Output:: PlayersValues 0 Jack Grealish £105.75m 1 Romelu Lukaku £103.50m 2 Jadon Sancho £76.50m.....

DESIGN – DATA SCRAPING SUBROUTINE & PSEUDO CODE ALGORITHMS

Pseudo Code For Data Scaping Values of Players

```
Import page parser

#headers function to ensure site doesn't reject request
        headers = {'User-Agent': }
        page = "URL of webpage"
#get page and headers
        page Tree = requests.get(page, headers)
# for all the players find values
        Players = find_all("span", {"class": "players"})
        Values = find_all("td", {"class": "values"})
#list players and values
        PlayersList = []
        ValuesList = []
        length = len(Players)
#create loop for all players in the list and append to dataframe
        for i in range(0,length):
            PlayersList.append(Players[i].text)
            ValuesList.append(Values[i].text)
        dataframe = pd.DataFrame({"Players":PlayersList[],"Values":ValuesList})
```

Sample Output

	index	Players	Values
0	0	D. de Gea	£16.20m
1	1	D. Henderson	£16.20m
2	2	T. Heaton	£900Th.
3	3	Lee Grant	£225Th.
4	4	R. Varane	£58.50m
5	5	H. Maguire	£43.20m
6	6	V. Lindelöf	£21.60m
7	7	E. Bailly	£7.20m
8	8	P. Jones	£3.60m
9	9	Luke Shaw	£37.80m
10	10	A. Telles	£16.20m
11	11	A. Wan-Bissaka	£34.20m
12	12	D. Dalot	£10.80m
13	13	S. McTominay	£31.50m
14	14	N. Matic	£5.40m
15	15	P. Pogba	£49.50m
16	16	D. van de Beek	£22.50m
17	17	Fred	£19.80m
18	18	B. Fernandes	£81.00m
19	19	J. Lingard	£18.00m
20	20	Juan Mata	£2.70m
21	21	M. Rashford	£76.50m
22	22	J. Sancho	£76.50m

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER: VARIABLES

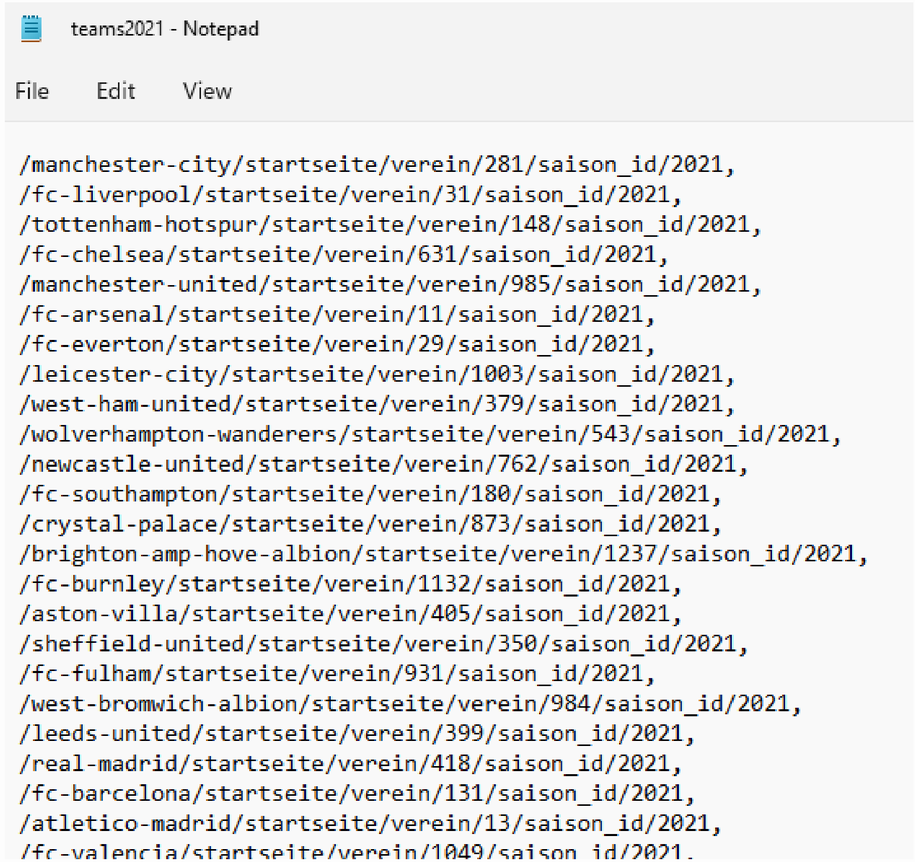
DESIGN – DATA SCRAPING CLASS DEFINITIONS & PSEUDO CODE

Class Definitions & Subroutines - Teams

The next subroutine will expand this concept for all the teams that I want. For each of the teams I have had to create a txt file with a unique URL. See the 'teams201.txt below. The code will loop through each of the teams to extract the player and values data for each team. This process is then repeated for previous seasons. I will use the class names of 'players' and 'values' that I defined earlier.

```
# read file with the URL of all teams
        teams=read_csv('teams2021.txt')
        length=len(teams)
#ensure that the site doesn't reject this request
Headers – ({user-agent:}
Df=Data.Frame()

#create loop for all teams
For x in range (length):
                page = 'url' + teams[x]
                pagetree = get(headers)
                        players = find_all, {"players"}
                values = find_all {values}
                PlayersList = []
                ValuesList = []
length len (Players)
Dataframe=df.append(dataframe1)
```



DESIGN – DATA ANALYTICS PSEUDO CODE

Introduction

Now that I have designed the data scraping algorithms, I will be able to develop the code to run on the websites and produce comprehensive.csv datasets for different years / seasons. These data sets will be used in the data analytics and the machine learning. The next stage in the design will be to develop the data analytics. This will generate very useful information that the user can use to compare player data. The output will be both in table format and in graphical format.

Data Analytics Pseudo Code

```
# Upload my master consolidated data file for each of the three seasons that I built through the data scrapping process
        data21 = read_csv(consolidated_data_2021)
        data20= read_csv(consolidated_data_2020)
        data19= read_csv(consolidated_data_2019)
#The next section is to plot the values of the most valuable players relative to their values
        values=data[“value”]
        avgvalue=statistics.mean(values)
        dataplot[‘player’]=[player names]
        ax.scatter(data[‘value’]
        ax.label(‘number of players’)
#The next analysis is to plot the mean values of the total number of players
        ax.axvline(avgvalue)
#Then plot the histogram of values
        ax.hist(values)
#the analysis to plot the log of values
        ax.hist(log_values)
```

DESIGN – DATA ANALYTICS PSEUDO CODE

Player Comparison Metrics Pseudo Code

The next part of the design for the data analysis section is to produce useful graphical analysis so that the user has an efficient way of comparing metrics of players. The graphical output will be used in the player report that the user will be able to access from the website.

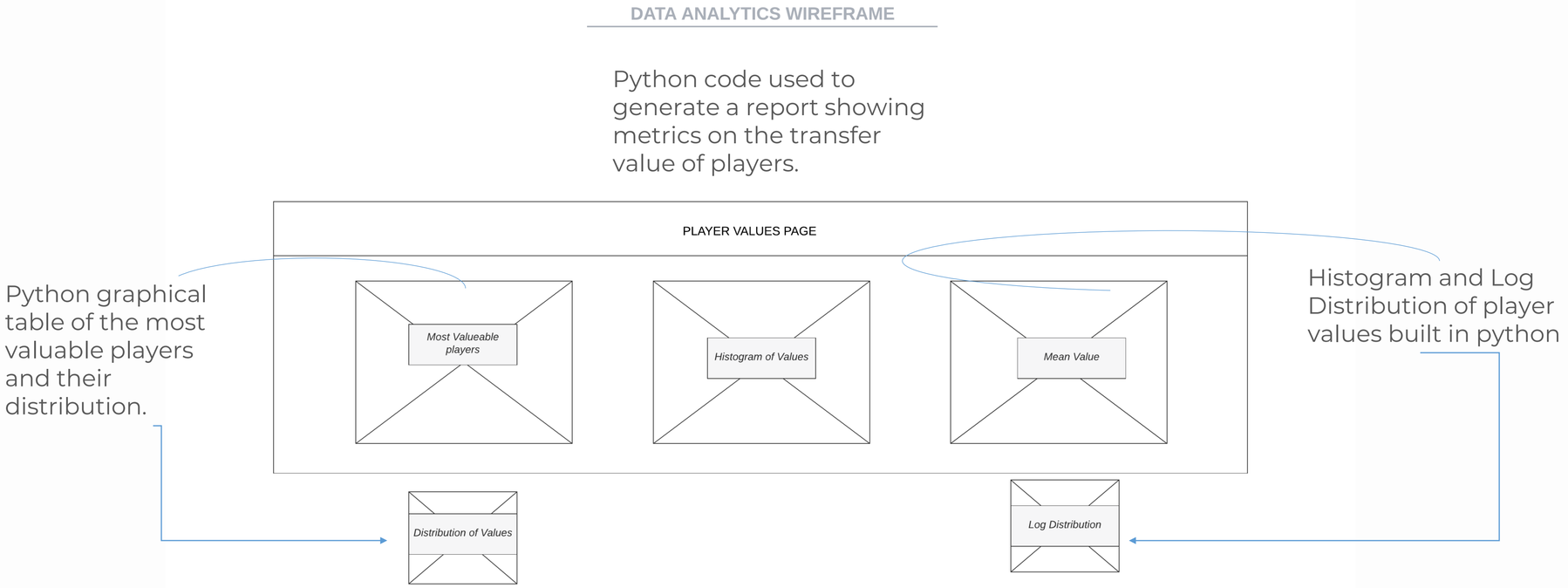
```
#import python graphical plotting
#compare technical metrics of player 1 and player 2 and transfer value
        Player_1 ={Pace, Shooting, Passing, Dribbling, Defending, Transfer_value'}
        Player_2 ={Pace, Shooting, Passing, Dribbling, Defending, Transfer_value'}
        data = dataframe
        attributes = list(data)
#indexing for selecting by position
        Values = dataframe.iloc
#angles for the chart
        angles = [pi for n in range]
#plot graph
        plot.subplot
        ax.plot(angles,values)
#fill graph plot
        ax.fill(angles, values, 'colour')
        Ax.set_title('Player_name')
#plot the graph
        plt.show
```

DESIGN - DATA ANALYTICS REPORT WIREFRAME

Data Analytics Report Wireframe

The goal of the data analytics section will be a professional looking report that the user can use to compare all the statistical attributes of a player.

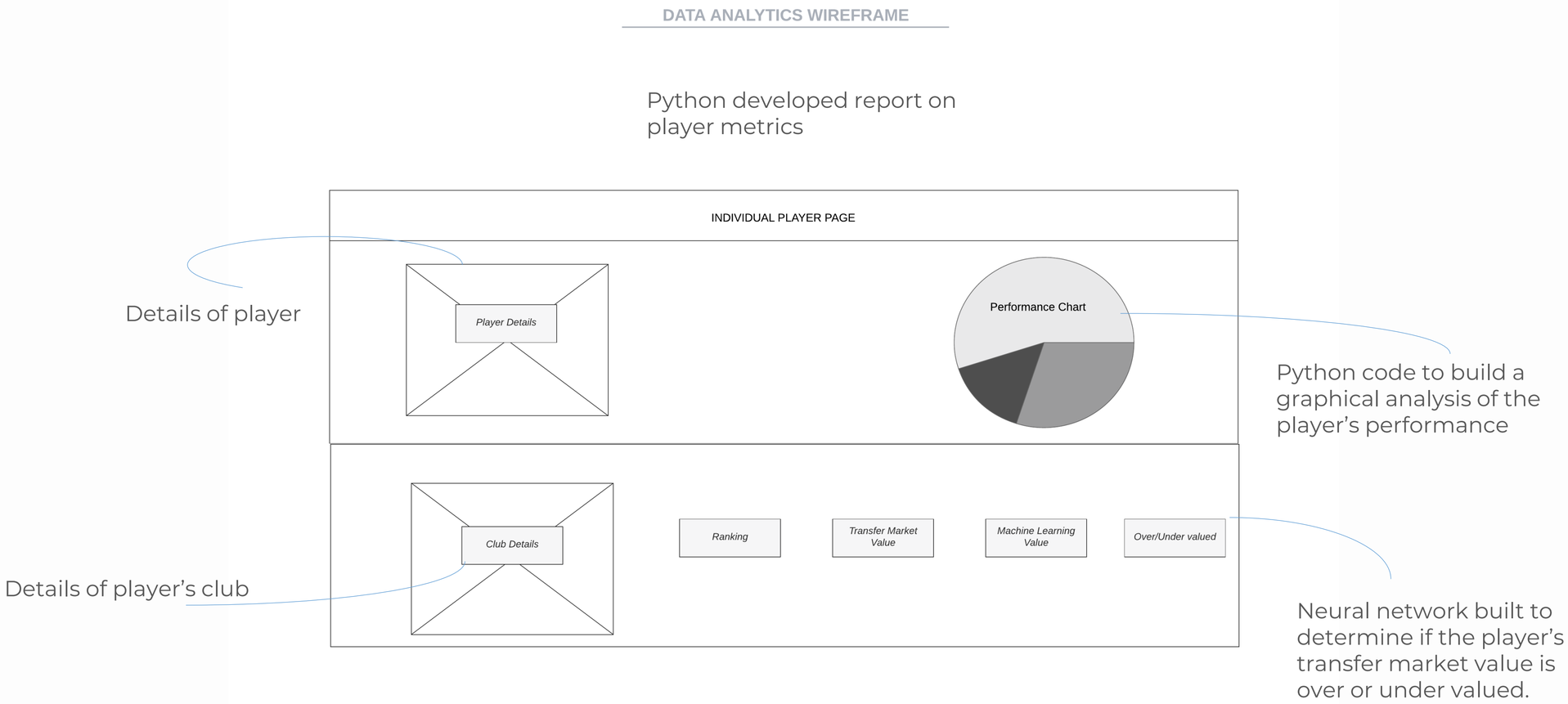
The first page of the report will the graphical chart showing the histogram and logarithm of the distribution of player values



DESIGN - DATA ANALYTICS WIREFRAME

Data Analytics Report Wireframe

The next page will show the details of the player. The second section will show the player's transfer market value and the machine learning algorithm's value. The user will then be able to note if the player is over or undervalued.



DESIGN – WEBSITE

AI.FOOTBALL

The design of the website will be professional and modern to appeal to the target user. Later in the design section I will design the website in more detail, but before doing so I need to consider the following:

Domain

The first part of the website design is to acquire an appropriate domain. After some considerable research I was able to purchase the domain: ai.football. This is an ideal URL for my website as the user will already know that it uses A.I and it is about football.

Hosting

In order to host the website, I need to setup a hosting package. I have chosen a large hosting provider as they have a 99.9% uptime guarantee as well as built-in resilience.

Security – SSL & 2FA

The domain users two factor authentication. This will ensure the security of the site as even if the password was compromised, they would not have the code to login. I shall also setup a https:// SSL certificate on the domain hosting for additional website security.

Development Tools

The site will require the following server-side development tools to be used:

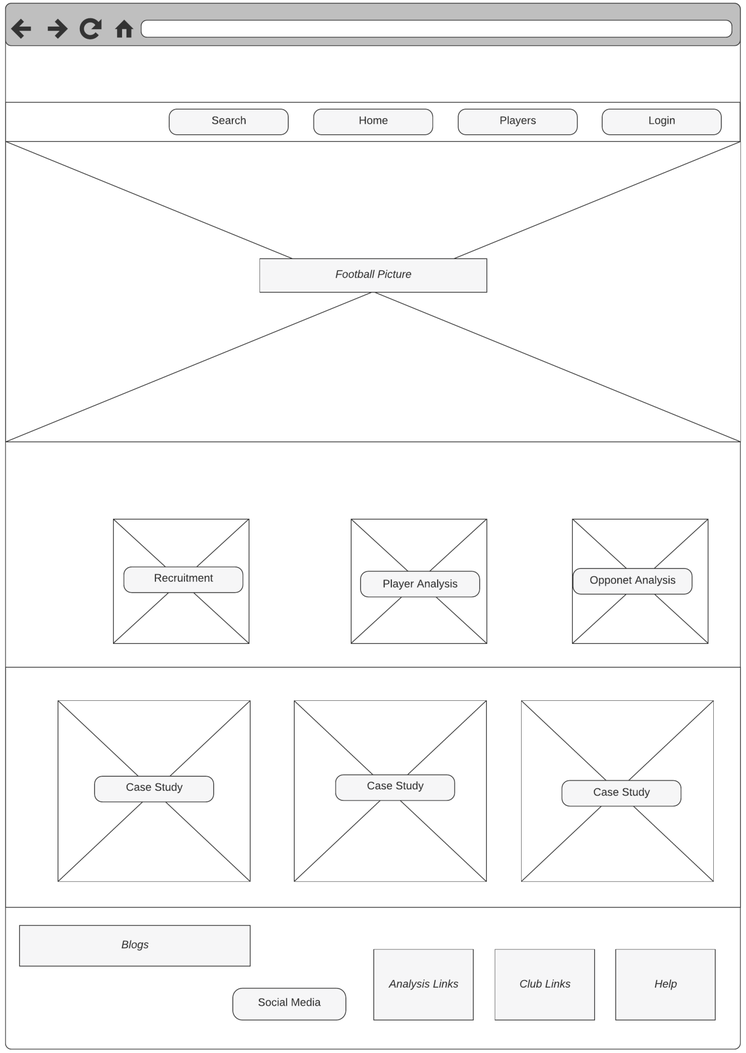
1. HTML
2. MySQL
3. phpMyAdmin

Logo & Images

I will design the ai.football logo using an online logo editor. The images used in the website have been purchased from the image website company www.shuttershock.com This ensures that the images I am using on my website are not in breach of any copyright.

DESIGN – WEBSITE WIREFRAME – FRONT PAGE

The website will have a modern front page with a striking picture



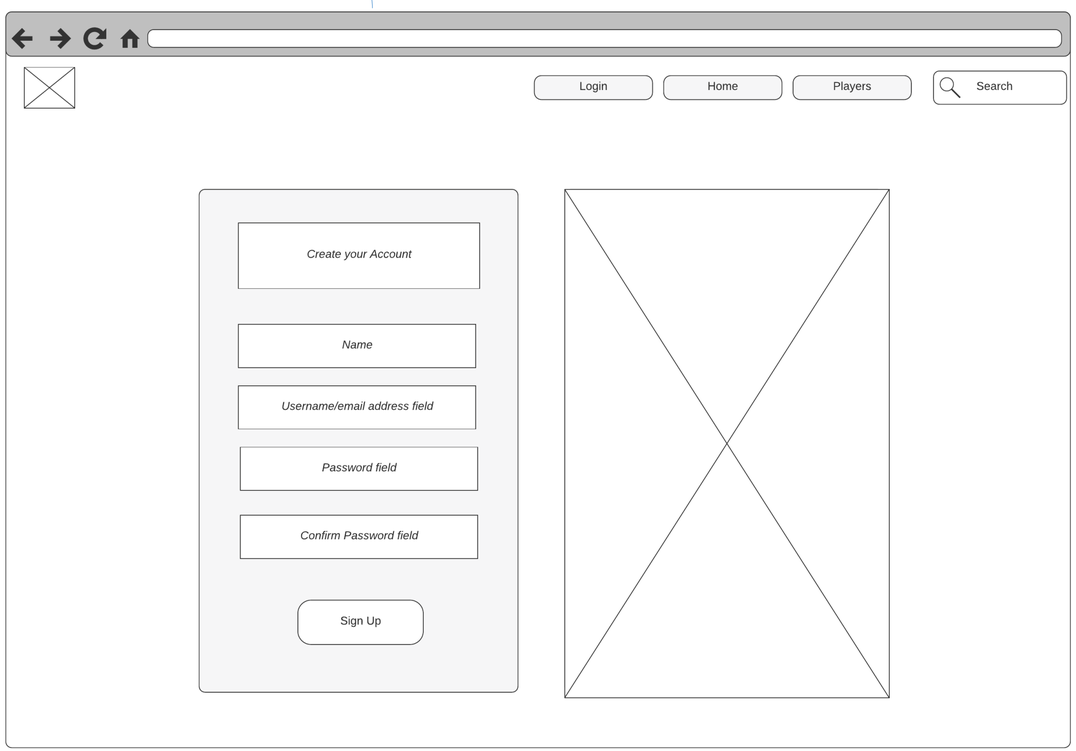
Links for easy navigation:

1. Search
2. Players
3. Login

DESIGN – WEBSITE WIREFRAME – REGISTER & LOGIN PAGES

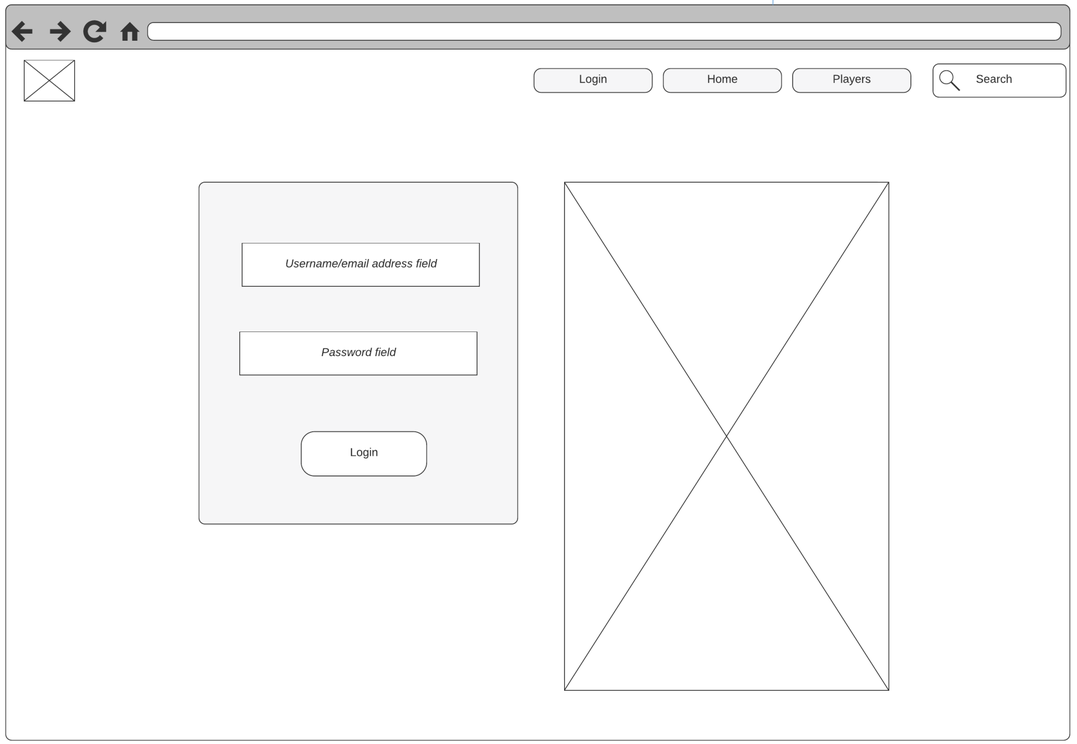
The registration page will allow the user to input their email and password

ai.football Register Page



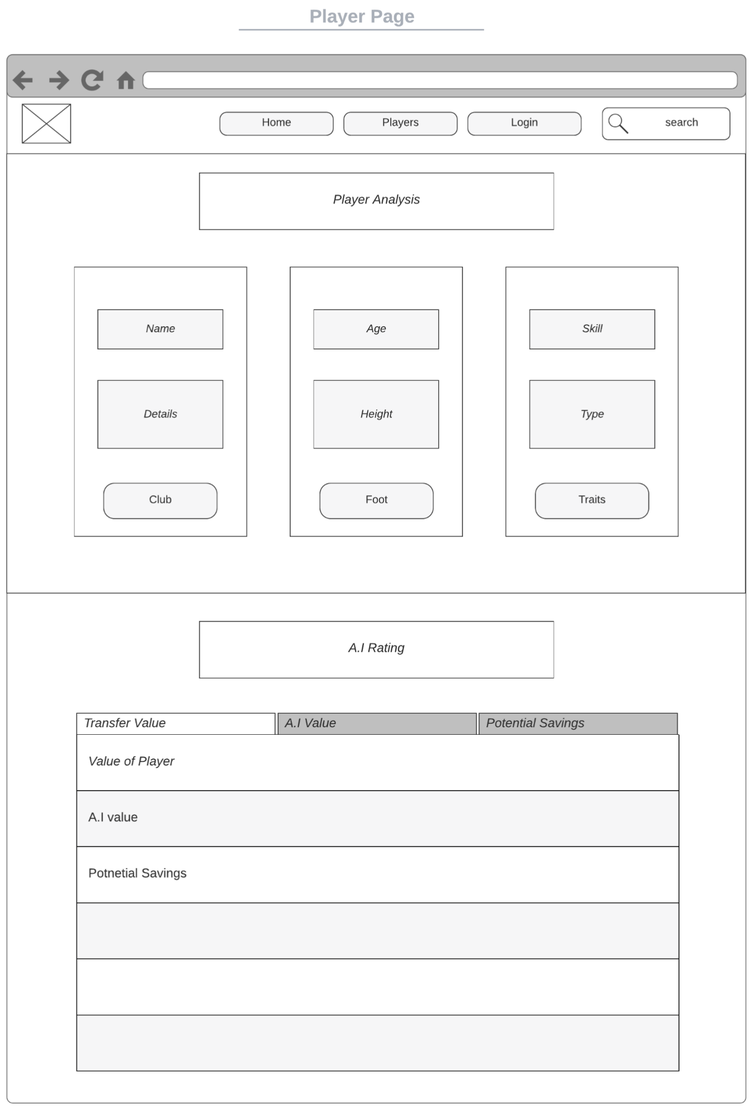
The login page will allow the user to login to access the ‘Players’ page

ai.football Login Page



DESIGN – WEBSITE WIREFRAME – PLAYER PAGES

The player page will have the player's key metrics



The player page will only be accessible once the user has registered and logged in

The A.I Section will have the transfer value of the player. It will also display the machine learning generated value of the player and the over / under valued metric according to the A.I algorithm

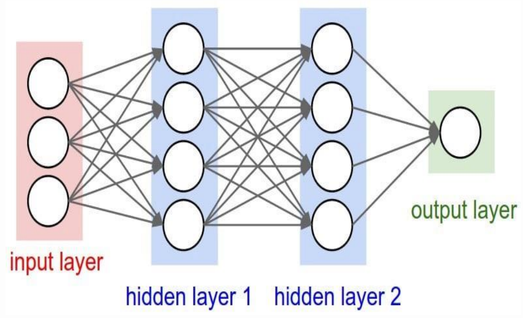
DESIGN

MACHINE LEARNING & NEURAL NETWORKS

DESIGN – NEURAL NETWORKS

Introduction

Neural networks are a set of algorithms, modelled after the human brain, that are designed to recognize patterns that are too complex for a human. They work through a set of interconnected layers. The first layer is the input layer, where the variables are inputted. The data is then communicated to the middle layers which are the “hidden layers”. This is where the actual processing is done via a system of weighted connections. The hidden layers then connect to an output layer where an output is produced.



A learning problem has a set of data and tries to predict properties of unknown data. Learning problems can be either:

1. Supervised learning

This is used when the data is in different data classes, and I want to learn from the data how to predict using a training set of data. If the output consists of more than one continuous variable, then this is called regression. An example of a regression problem would be the prediction using data such as goals scored, height, age etc.

2. Unsupervised learning

Used to find groups of similar examples in a set of data. This type of neural network learning is therefore not appropriate for what I require and will be using a supervised learning algorithm.

DESIGN – NEURAL NETWORKS & REGRESSION

Application

One of objectives identified in the objectives section was to design a system that would identify what the primary factors are, to determine the valuation of a player and is there a way to use machine learning to identify whether a player is over and under valued'?

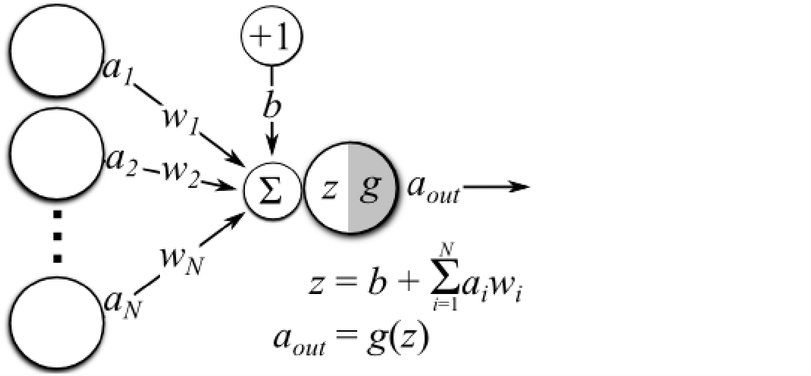
Neural Network

A neural network can be used to predict an output given a series of inputs. When I first researched about neural networks it seemed to be a "mystical" concept. However, when I researched further, I then appreciated that a neural network uses a mathematical concept know as regression. The main idea behind regression is to apply weights to each statistic, with more useful and recent statistics having a bigger weight and outliers not having any weight. The output is then a function of a sum of the product of weights and inputs. Mathematically this can be summarized as follows:

$$z = x_1 * w_1 + x_2 * w_2 + \ldots\ldots + x_n * w_n + b * 1$$

$$\hat{y} = a_{out} = sigmoid(z)$$

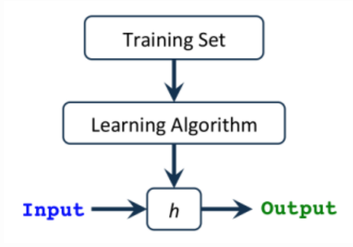
$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$



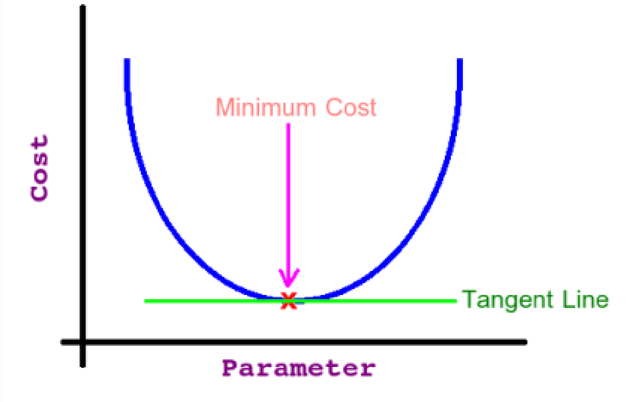
DESIGN – NEURAL NETWORKS & COST FUNCTION

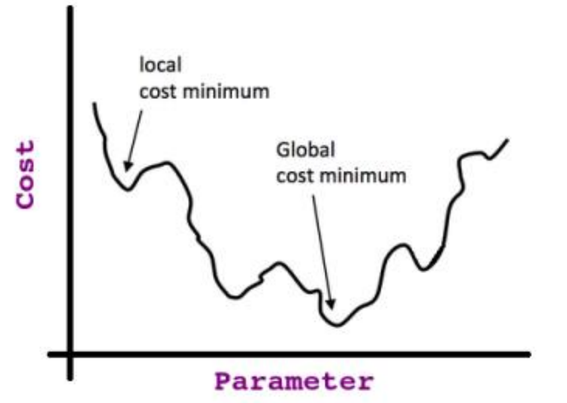
Least Squares & Linear Regression

Regression is one of the most used algorithms for machine learning classifications.. A regression problem is one in which I am trying to predict a value of a continuous variable, like predicting the value of a player. In my system I want to take a hypothesis h(x), which takes an input and gives me an estimated output value. This hypothesis can be as simple as a one-variable linear equation up to a very complicated and long multivariate equation.



My task is to find the best parameters (or weights) that give me the least error in predicting the output. The function that calculates this error is a Cost(or Loss) Function and my goal is to minimize the error in order to get the best predicted output. The relation between the parameter value and its effect on the cost function (or the error) which looks like a bell curve. So, if I start at any point in that curve and keep taking the derivative (tangent line) of each point I will end up at what so called the Global Optima as shown in this image: 'the bell curve'. If I take the partial derivative at the minimum cost point (i.e. global optima) I find the slope of the tangent line = 0 (then we know that we reached our target). That's valid only if I have a Convex Cost Function, but if I don't, we may end up stuck at what is called Local Optima as illustrated in this non-convex function.





DESIGN – NEURAL NETWORKS & COST FUNCTION

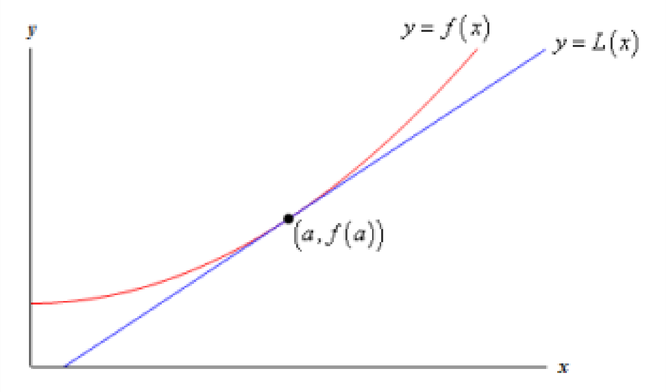
Cost Function

Machine learning uses derivatives in optimization problems. Derivatives are used to decide whether to increase or decrease the weights or decrease an objective function. The solver then tries to find the parameter weights that minimize a cost function. This is the mathematical concept where we have a function with n variables, then the gradient is the length-n vector that defines the direction in which the cost is increasing most rapidly. So, in gradient descent we follow the negative of the gradient to the point where the cost is a minimum. In machine learning the costs' function is a function to which we are applying the gradient descent algorithm.

Linear Approximation

Given a function, f(x), we can find its tangent at x=a. The equation of the tangent line L(x) is: L(x)=f(a)+f'(a)(x−a).

In the following graph of a function and its tangent line. From this graph x=a, the tangent line and the function have nearly the same graph. On occasion, we will use the tangent line as an approximation to the function f(x) In these cases, we call the tangent line the "Linear Approximation" to the function at x=a



DESIGN – NEURAL NETWORKS & WEIGHTS

Weights

As I introduced the concept of weights, I wanted to note how I will design this concept in my machine learning algorithm. The weight is the parameter within a neural network that transforms input data within the network's hidden layers. The neural network that I will use has a series of nodes. Within each node is a set of inputs, weight, and a bias value. As an input enters the node, it gets multiplied by a weight value and the output is either observed or passed to the next layer in the neural network. These weights are used to assign more importance to a particular dataset used for the learning.

In the design of my system, I will apply the following weights for each of the three football seasons I have the data for. Weights can be between 0 and 1. The weight with the highest value of 1 will apply to the most recent season. The two previous seasons will have less weights applied as this reflects that although the data is useful, it is older than the most recent season and understandably carries less relevance.

1. Season 2021: Weight = 1
2. Season 2020: Weight = 0.8
3. Season 2019: Weight = 0.7

Sklearn

There are many different neural networks that I could use, but I have chosen the ‘sklearn’. It is one of the most popular neural networks. Solving logistic regression is an example of an optimization problem. The sklearn has several solver algorithms that I can use. I have chosen the most used solver called lbfgs which stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno.

DESIGN – LEAST SQUARE & PYTHON STATISTICAL ANALYSIS

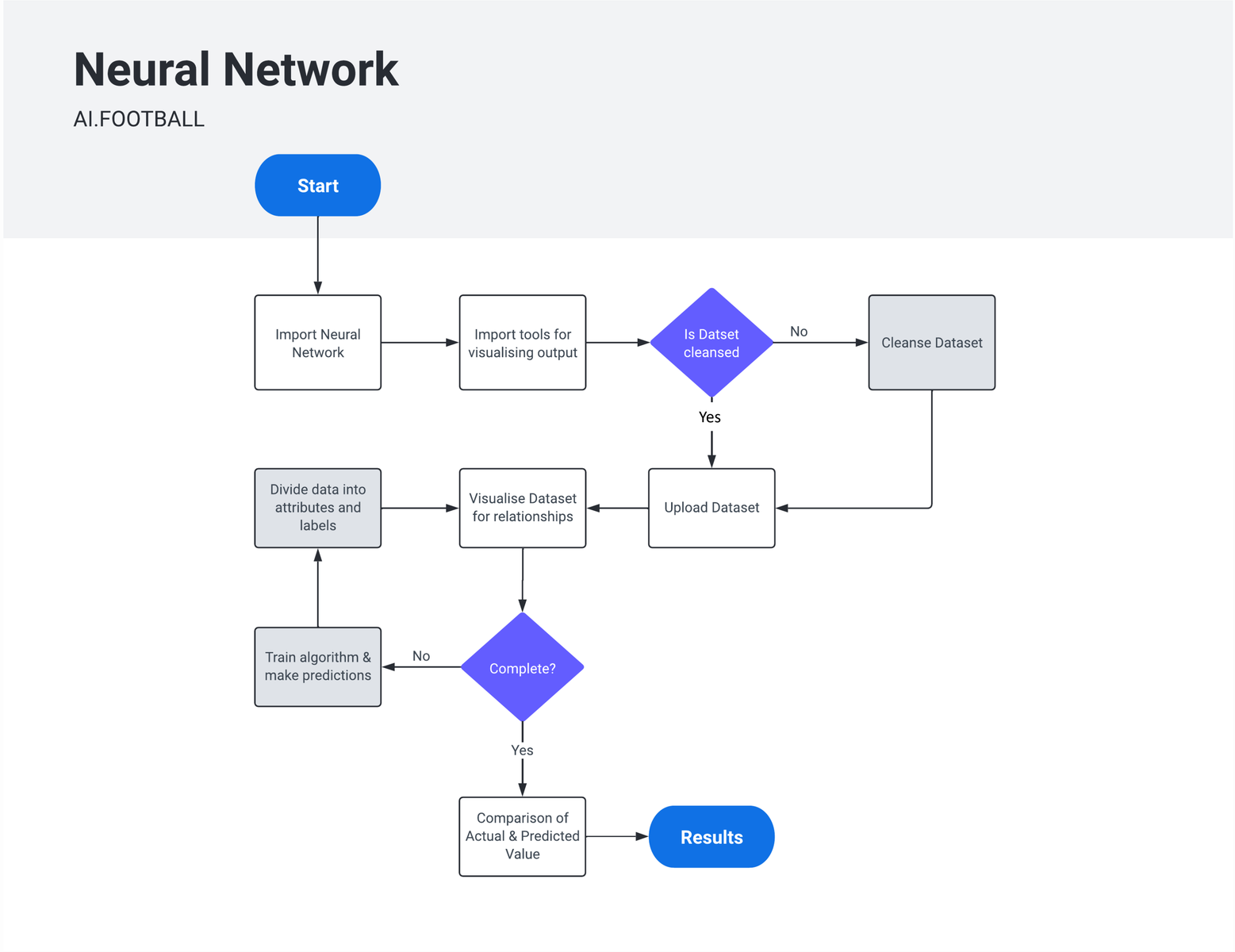
Statistical Algorithm The statistical method I have used to model the different player positions is a reweighted least square model. The value of a player is determined by several factors including their on-pitch statistics, their physical attributes such as height and age, but also what league the player plays in.

Python Statistical Libraries

I have tried to limit the use of libraries, but as there is a lot of statistical analysis I will use the following python statistical libraries for the mathematical computation.

(i) matplotlib - for the graphical plotting
(ii) seaborn - for data visualization
(iii) Statsmodels - for linear regression
(iv) Sklearn – neural network

DESIGN – NEURAL NETWORK ALGORITHM



DESIGN – NEURAL NETWORK STRUCTURS & CLASS VARIABLES

Data Structures

The next part of the design will be to define the primary data structures that will be used in the machine learning code. The better the on-pitch statistics the player has, the higher their value is likely to be. However, there are several other factors that I will test to see if they contribute to the overall value. I have identified the following variables that will be used for the neural network input which will be tested in the Testing section to see if the neural network can predict the value of a player

Class Variables

1. Age {integer}: this variable will be put into the neural network code and tested to see if it's statistically significant. My design hypothesis is that the older the player is, on average the lower the value.
2. Player position {varchar}: to test the hypothesis that the position a player plays is significant in assessing the value of a player.
3. League{ varchar}: - to assess if the league a player plays in is important. The hypothesis being that the Premier league has higher valued players than fourth division.
4. Team performance {varchar} - evaluate is the team a player plays in. My design hypothesis is that here is a higher probability that the better teammates the player has, the higher the potential value of a player. I have therefore defined the statistically significant variables which measure the team's overall ability as follows:
    - CL –{integer} binary measuring if the player's team played in Champions League following season,
    - Pts – {integer} player's team points
    - xG – {integer} player's goals
    - xGA – {integer} player's team goals against

DESIGN – DATA STRUCTURES

Class Variables

The input variables into the neural network will be split into sections for each of different player positions – goalkeepers, midfielders and forwards as each position has attributes which are a factor in determining the player value.

Goalkeepers

For goalkeepers, the following variables will be used for the training neural network:

- clean_sheetsm - goalkeeper's clean sheets per a minute played
- psnpxg_per_shot_on_target_against - Post-shot non-penalty expected goals per shot on target against the player playing as a goalkeeper
- passes_pct_launched_gk - % of passes that were launched

Midfielders

For midfielders, these are statistically significant variables that will used:

- goals – number of goals player scored
- xg_xa_per90 – sum of expected goals and expected assists per 90 minutes
- passes_completed_short - passes completed between 5 and 15 yards
- passes_into_final_third - passes into attacking third of football pitch
- carry_distance - total distance the player moved with the ball by his feet
- tackles_won - number of tackles won

Forwards

- goals – number of goals player scored
- xg_xa_per90 – sum of expected goals and expected assists per 90 minutes
- passes_into_final_third - passes into attacking third of football pitch
- touches_att_pen_area - touches in competitor's team penalty area
- dribbles_completed – number of player's successful dribbles

DESIGN – LINEAR REGRESSION GOALKEEPERS PSEUDO CODE

Pseudo Code - Goalkeepers

The neural network algorithm and pseudo code for the training the data set of players who are designated goalkeepers.

```
Import sklearn

#upload the data csv file with the player data
        Data = read(consolidated_datacsv_files)

#select only the goalkeepers and remove the outlier data for those goalkeepers that would affect the machine learning algorithm,
        dataGK = goalkeepers
#train neural network using the factors that are statistically significant in a goalkeeper's value
        trainGK =  (dataGK)
        Model=(age+League=Premier_League)
        Model=(goals+pts_XGApsnpxg_per_shot_on_target_against+ passes_pct_launched_gk
#results and create a regression
        Results=modelGK.fit
        modelGK.regression
        finalGK = regression.model.OLSResults
#display output
        finalGK.summary
```

DESIGN – LINEAR REGRESSION DEFENDERS PSEUDO CODE

Pseudo Code - Defenders

The neural network algorithm and pseudo code for the training the data set of players who are designated defenders.

```
Import neural network
#upload the data csv file with the player data
        Data = read(consolidated_datacsv_files)

#select only the defenders and remove the outlier data for those defenders that would affect the machine learning algorithm,
        dataDEF = defenders
#train neural network using the factors that are statistically significant in a midfielder's value
        trainDEF =  (dataDEF)
        Model=(age+League=Premier_League)
        Model=(goals++xg_xa_per90+ 'passes_ground+touches_att_pen_area+touches_def_pen_area+aerials_won_pct')
#results and create a regression
        Results=modelFWD.fit
        modelDEF.regression
        finalDEF = regression.model.OLSResults
#display output
        finalDEF.summary
```

DESIGN – LINEAR REGRESSION MIDFIELDERS PSEUDO CODE

Pseudo Code - Midfielders

The neural network algorithm and pseudo code for the training the data set of players who are designated midfielders.

```
Import neural network
#upload the data csv file with the player data
        Data = read(consolidated_datacsv_files)

#select only the midfielders and remove the outlier data for those midfielders that would affect the machine learning algorithm,
        dataMD = midfielders
#train neural network using the factors that are statistically significant in a midfielder's value
        trainMD =  (dataMD)
        Model=(age+League=Premier_League)
        Model=(goals+passes_completed+dribbles+ passes_final_third+touches_penalty_area)
#results and create a regression
        Results=modelFWD.fit
        modelFWD.regression
        finalFWD = regression.model.OLSResults
#display output
        finalFWD.summary
```

DESIGN – LINEAR REGRESSION FORWARDS PSEUDO CODE

Pseudo Code - Forwards

The neural network algorithm and pseudo code for the training the data set of players who are designated forwards.

```
Import neural network
#upload the data csv file with the player data
        Data = read(consolidated_datacsv_files)

#select only the forwards and remove the outlier data for those forwards that would affect the machine learning algorithm,
        dataFWD = forwards
#train neural network using the factors that are statistically significant in a forward's value
        trainFWD =  (dataFWD)
        Model=(age+League=Premier_League)
        Model=(goals+passes_completed+pts psnpxg_per_shot_on_target_against+ passes_final_third)
#results and create a regression
        Results=modelMD.fit
        modelMD.regression
        finalMD = regression.model.OLSResults
#display output
        finalMD.summary
```

DESIGN – NEURAL NETWORK MLP CLASSIFER

MLP Classifier

The Class MLPClassifier is a multi-layer perceptron (MLP) algorithm that trains using Backpropagation. MLP trains on two arrays: array X of size (n_samples, n_features), which holds the training data as a vector and array y of size (n_samples,), which holds the target values which are the class labels for the training samples.

MLP trains using a form of gradient descent and the gradients are calculated using Backpropagation. It gives a vector of probability estimates. One of the advantages of this neural network is that it supports multi-label classification in which a sample can belong to more than one class. For each class, the output passes through the logistic function.

The MLPClassifer algorithm is as follows:

```
from sklearn.neural_network import MLPClassifier
#define arrays
        X = [[0], [1]
        y = [0, 1]
# use the solver function
        clf = MLPClassifier(solver)
        hidden_layer_sizes=(),)
        clf.fit(X, y)

#After fitting (training), the model can predict for new samples:
        clf.predict([], [])
        array([])
```

DESIGN – NEURAL NETWORK PREDICTIONS PSEUDO CODE

Machine Learning Predictive Outcomes Pseudo Code

Now that I have designed the code for each of the player positions, I will need to run the neural network predictive model. This algorithm takes the player code and splits the data into a learning and training set. It then applies the MLP classifier and solver to determine how many of the predictions are correct.

Once the code for each of the positions has been developed, I need to import this code to run the neural network algorithm

```
        Import My_GoalkeepersGK_code
        data=GK.dataGK
#create table of over or under valued by taking player value from our calculated predsOLS value
        data['Over/undervalued']=np.where(data['value']-data['predsOLS']>0, 'Overvalued', 'Undervalued')

#splitting data into test and training dataset
        yGK = data['Over/undervalued']
#factors that influence goalkeeper values
        XGK_train, XGK_test, yGK_train, yGK_test = train_test_split(XGK,yGK,)

#applying MLPClassifier algorithm and solver to predict the % of correct predictions
        net=MLPClassifier()
        netsolver='lbfgs'
        net.fit(XGK_test,yGK_test)
        ytest_predict=net.predict(XGK)
        outcome=len(trueorfalse[trueorfalse=='T'])
        print('There are ' + outcome + '%' + ' correct classifications')
```

DESIGN

AI.FOOTBALL

DESIGN – AI.FOOTBALL

Website Structure

The website has the following file structure:

1. Front Page: ‘index.php”
2. Registration Page: ‘register.php’
3. Login Page: ‘login.php’
4. Players Page: ‘players.php’
5. Player Details: ‘player_details.php’

Website Development Coding Languages

I have chosen to use the following development tools as they are widely used opensource programs for website and database development

1. HTML – this is for the design of the webpages and includes the images for the site
2. PHPMyAdmin –is a free, open-source administration tool for MySQL. It has become one of the most popular MySQL administration tools, especially for web hosting services
3. MySQL – this is a relationship database management system. MYSQL is the most popular database server. By using the PHPmyadmin I can administer MYSQL. I have used this as it will enable me to view the databases and tables.

MySQL Data Structures

The design of ai_football MySQL database will have two table structures:

1. Players – which is used to contain the data on the players
2. Users – which contains the details of the users who register on the website

The following website schematic provides the detail of the interaction of the pages and the database.

DESIGN – AI.FOOTBALL SECURITY

SSL Certificate

SSL certificates create an encrypted connection and establish trust. One of the most important components of online business is creating a trusted environment where potential customers feel confident in making purchases. SSL certificates create a foundation of trust by establishing a secure connection. SSL certificates have a key pair: a public and a private key. These keys work together to establish an encrypted connection. The certificate also contains what is called the “subject,” which is the identity of the certificate/website owner.

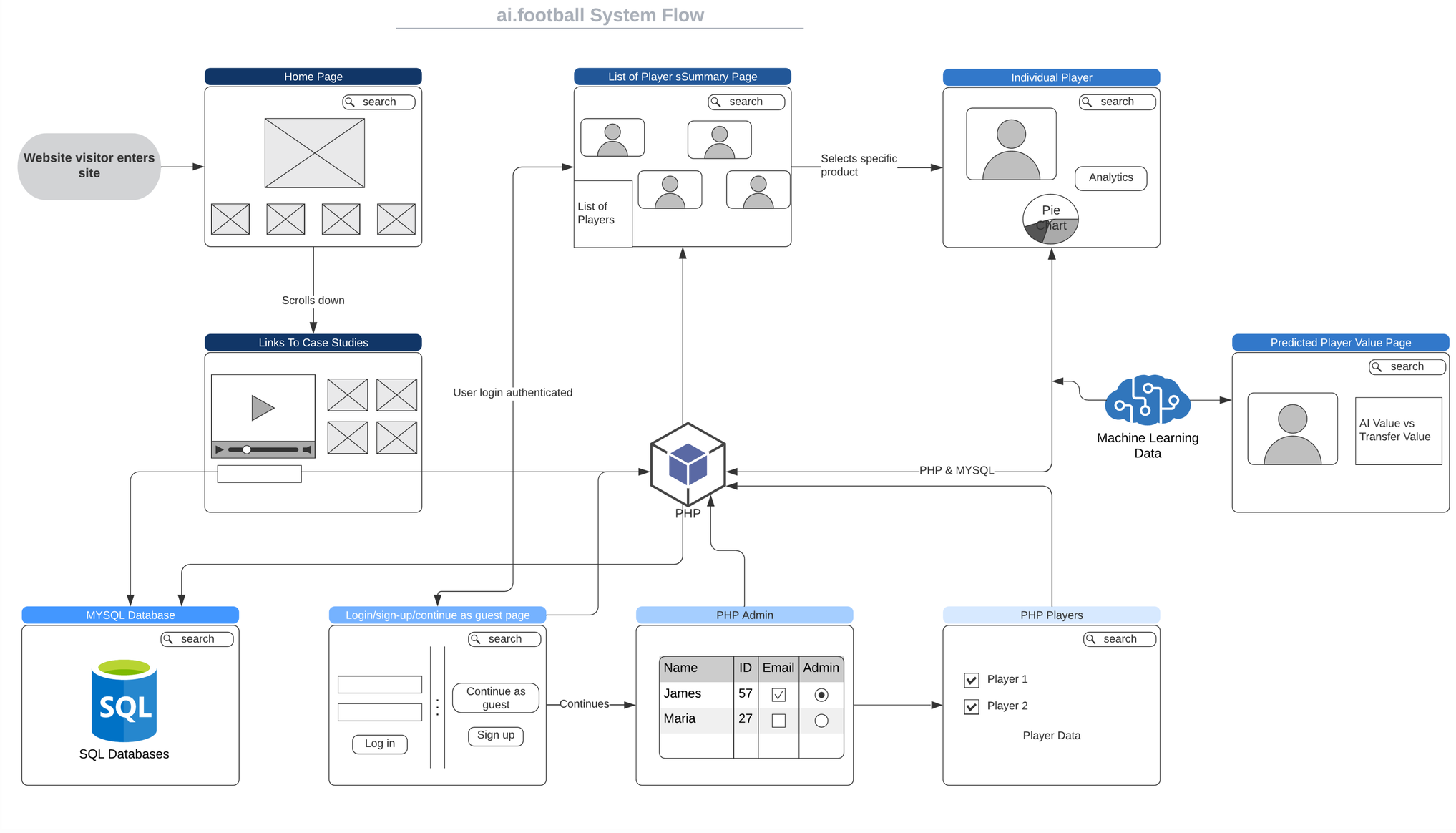
To get a certificate, I need to create a Certificate Signing Request (CSR) on my server. This process creates a private key and public key on my server. The CSR data file that me send to the SSL Certificate issuer (called a Certificate Authority or CA) which contains the public key. The CA uses the CSR data file to create a data structure to match my private key without compromising the key itself. The CA never sees the private key.

Once I have the SSL certificate, I will install it on my server. I need to also install an intermediate certificate that establishes the credibility of my SSL certificate by tying it to my CA’s root certificate.

The most important part of an SSL certificate is that it is digitally signed by a trusted CA. Anyone can create a certificate, but browsers only trust certificates that come from an organization on their list of trusted CAs. Browsers come with a pre-installed list of trusted CAs, known as the Trusted Root CA store. In order to be added to the Trusted Root CA store and thus become a Certificate Authority, a company must comply with and be audited against security and authentication standards established by the browsers.

An SSL certificate issued by a CA to an organization and its domain verifies that a trusted third party has authenticated that organization’s identity. Since the browser trusts the CA, the browser now trusts that organization’s identity too. The browser lets the user know that the website is secure, and my users can feel safe browsing the site and even entering their confidential information

DESIGN – AI.FOOTBALL SCHEMATIC FLOWCHART



DESIGN - AI.FOOTBALL DATABASE STRUCTURE

MySQL Table Structure - Players

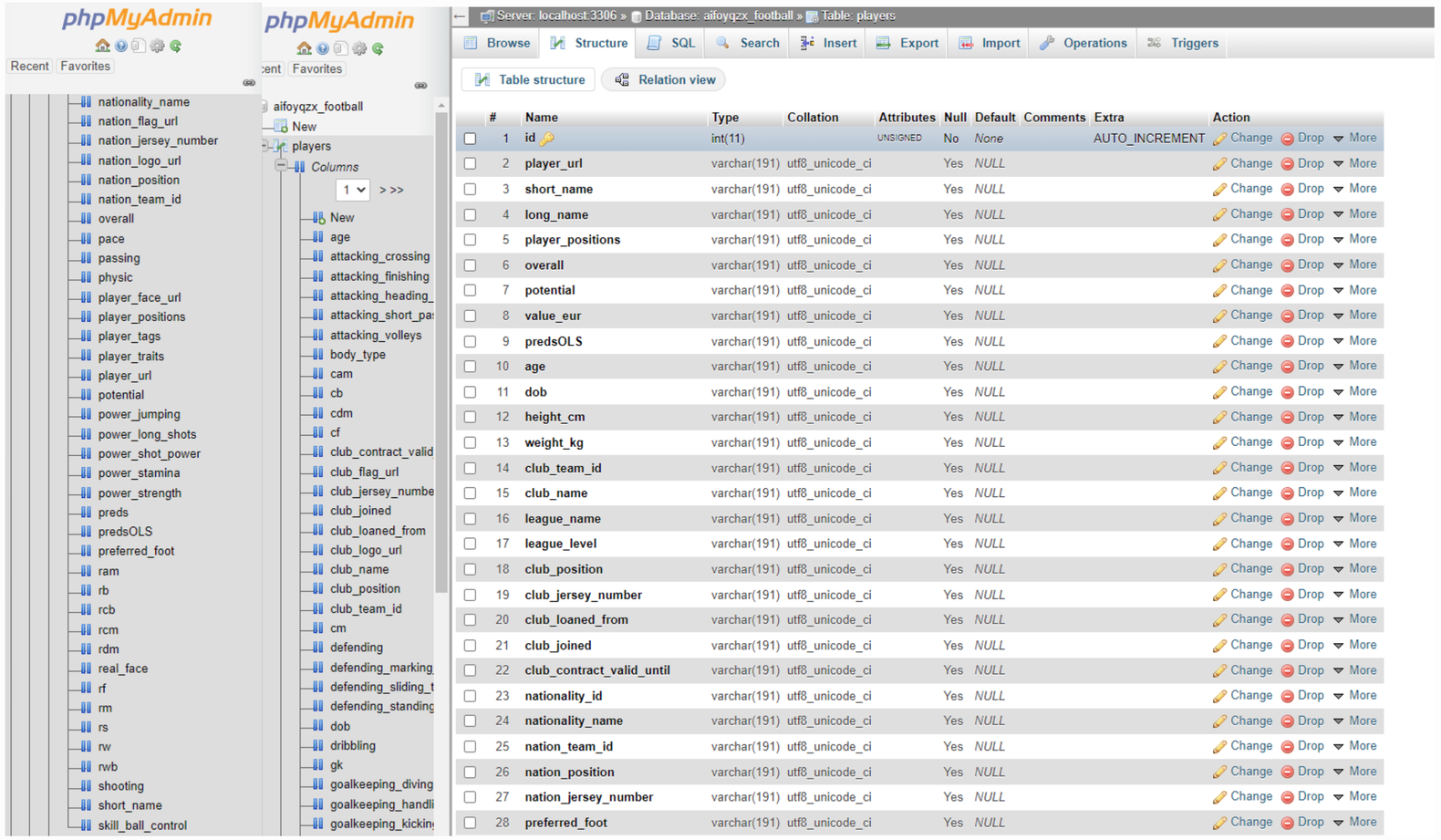
The MySQL table structure for the user login will have the fields containing the player attributes. Each of the these will be defined as ‘varchar’ type except for ID. The ID is an integer which will be the unique identifier for each player (as two players could have the same name). The following are some of the fields and their attributes.

```
-- Host: localhost:3306
-- Database: `ai_football`
-- Table structure for table `players`
CREATE TABLE `players` (
  `id`                  int(11)          (primary key)
  `short_name`          varchar(191)
`long_name              varchar(191)
`overall`               varchar(191)
`potential`             varchar(191)
`value`                 varchar(191)
`predsOLS`              varchar(191)     predsOLS is the field with the output value from the neural network training
`age`                   varchar(191)
`dob`                   varchar(191)
`height_cm`             varchar(191)
`weight_kg`             varchar(191)
`club_team_id`          varchar(191)
`club_name`             varchar(191)
`league_name`           varchar(191)
```

DESIGN – AI.FOOTBALL DATABASE STRUCTURE

MySQL Table Structure - Players

The following is the design of the MySQL database which I will manage using phpMyAdmin.

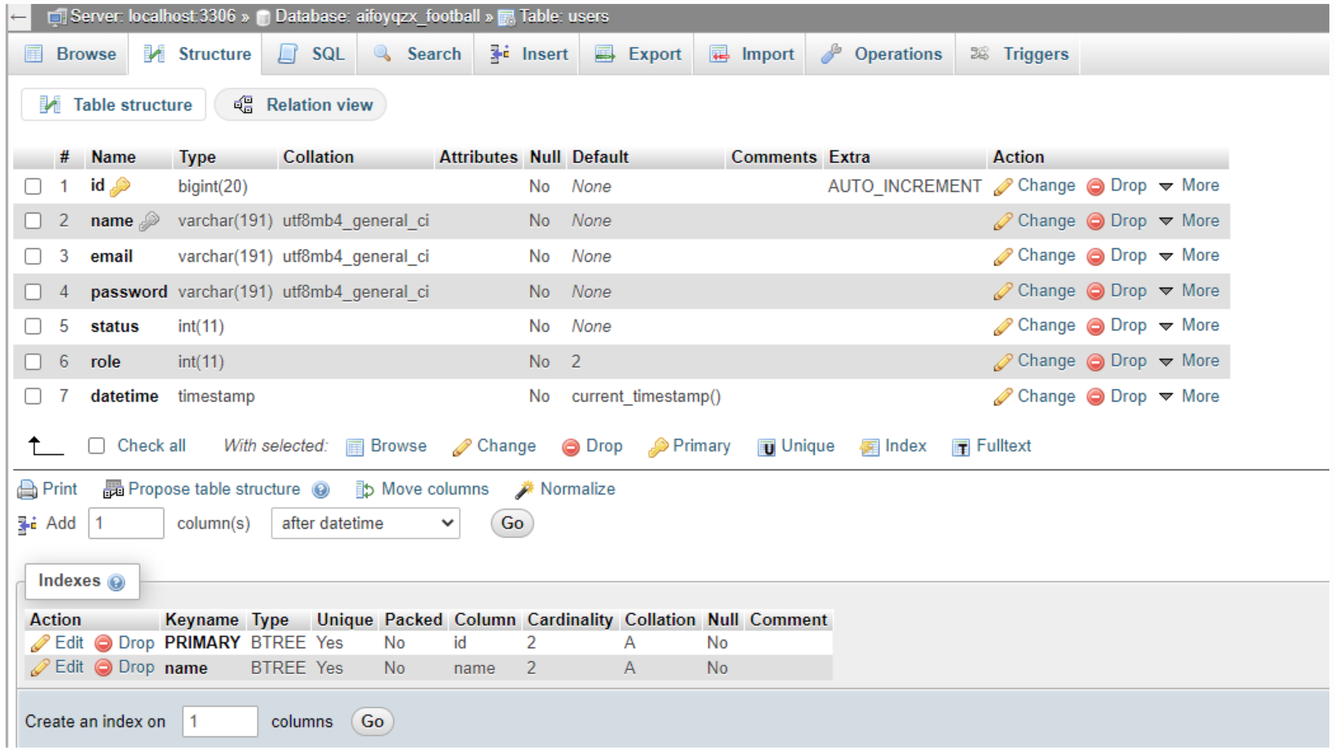


DESIGN – AI.FOOTBALL USER REGISTRATION

Database Table Structure – Register User

The php table structure for the user login will have the following input fields:

Id	Integer (primary key)
Name	Varchar
Email	Varchar
Password	Varchar (encrypted)
Status	Integer
Role	Integer
Datetime	timestamp



DESIGN – AI.FOOTBALL FRONT & HEADER PAGE

Index.php

The front page of the ai.football is index.php. This links to headers.html which contains the menu options. The pseudo code is:

```
 <head>
   <title>ai.football</title>
<body class="home page">
<php include('header.php');
                    <h1 class="hero-slide__title">Machine learning to build better teams</h1>
```

Header.html

```
<?php
<header class="header">
   <div class="header__top">
                    ul id=main-menu="menu">
                    href="<Home>
<!-- check if user is logged in else register-->
                    php if (session->isLoggedIn())
                    href="<login.php">Login</a>
                     <li id="menu-item"
                            href="<register.php">Sign Up</a>
                  <?php } else {
                     <li id="menu-item"
                        class="menu-item href="#">Welcome, <?php loggedInUserObj->name; >
<a class="nav-trigger nav-trigger--new" id="nav-trigger--new" >
```

DESIGN – AI.FOOTBALL USER REGISTERATION

Register.php

The registration page is register.php. This pseudo code stores the name, email and password. This is then stored in the ‘users’ table with a unique ID for every user. The password is encrypted so that the user’s password cannot be compromised.

```
<?php require_once("initialize.php");
// user information
   $name            = $_POST['name'];
   $email          = $_POST['email'];
   $password       = $_POST['password'];
   $confirmPassword = $_POST['confirm_password’];
 $newlyCreatedUserId = $userCreated;

$session->message("Thank you for registration. kindly login below.");
          header("location:login.php");
          exit;
```

DESIGN – AI.FOOTBALL USER LOGIN

Login.php

The registration page is login.php. The design requires the user to enter their details and then authenticates if correct. If the details are not correct, then there is an error message displayed informing the user that the combination is incorrect.

```
<?php require_once("initialize.php");
 {
       //username or email
       //check if the form is submitted
       if (isset($_POST['submit']))
          $log = trim($_POST['email']);
          $password = trim($_POST['password']);
                        $user_obj = User::authenticate($log, $password);
       if ($user_obj)
          if ($user_obj->status!= 0)
             $session->login($user_obj);
             header("location:login.php");
             exit;
} else {
          $arr_errors[] = "Email/password combination is incorrect.";
       }
    } else {
       $arr_errors[] = "Please enter email and password.";
```

DESIGN – AI.FOOTBALL PLAYERS PAGE

Players.php

Once the user has logged in, they will be able to select from the menu option the ‘Players’ page. This will load the player details and their key metrics.

```
<?php require_once("initialize.php");
global $database;
if (!$session->isLoggedIn()) "login.php");
$player = Player:($_GET['id']);
<h1 class="hero-slide__title">Player Analysis</h1>
                    <p>Here are the statistics of <?php echo $player->long_name;
                     <h2> <?php echo $player->long_name?></h2>
                    <p>ai.football uses the power of artificial intelligence to understand what drives player valuation.</p>
                    Date of birth: <?php echo $player->dob; ?>
                    Age: <?php echo $player->age; ?>
                    Rating: <?php echo $player->rating ?>
                    Height: <?php echo $player->height_cm; ?>
                    Preferred Foot: <?php echo $player->preferred_foot; ?>
                    Skills: <?php echo $player->skills; ?>
</div>
```

DESIGN – AI.FOOTBALL PLAYER DETAILS & A.I VALUATION

Players-Details.php

The player details page is where the user can access the A.I analysis. This user searches for a particular player and it returns the player details, the player's transfer value and the machine learning derived valuation. The system will also display the difference between the transfer value and the A.I value to give the user the value of how much that player is over or undervalued..

```
<div class="table__label">Overall<>
 <div class="table__title">A.I Rating<>
<div class="table__label">Player<>
 <div class="table__title">Transfer Value<>
<php echo $player->value>
```

the field 'predsOLS' is the output from the neural network to predict the value of the player

```
div class="table__price"><span><?php echo $player->predsOLS;?>
<div class="table__label">Player</div>
```

to calculate whether the player is over or undervalued, the 'value' field is taken away from the 'predsOLS' value to give 'potential savings'

```
<div class="table__title">Potential Savings</div>
<div class="table__price"><span><?php echo ($player->value - $player->predsOLS);
      <div class="table__time">ai.football<>
```

DESIGN – OVERALL DESIGN FILE STRUCTURE

File Structure

The overall design structure of the data and code for the system is shown in the following tree diagram.

TECHNICAL SOLUTION

THE CODE

CODE - OVERVIEW

Documentation

The code is listed in a separate coding document. A summary of the each of the sections is as follows:

1. Data Scraping
    (i) Scraping _code_single_team.py
    (ii) Scarping_code_teams.py

2. Data Analytics
    (i) Data_analytics.py
    (ii) Player_comparison.py

3. Machine Learning
    (i) Player_comparison.py
    (ii) Data_analytics.py
    (iii) ML_predictive_values.py
    (iv) Neural_network.py

4. Website
    (i) Inatlize.php
    (ii) Index.php
    (iii) Headers.php
    (iv) Login.php
    (v) Register.php
    (vi) Players.php
    (vii) Player_details.php

TESTING

DEVELOPMENT & TESTING

TESTING

Methodology

Throughout the development of the system, I will be testing it to make sure that it works as I set out in the Analysis section. This is important as it will highlight any bugs, areas where the system doesn't work fully or areas that need to be developed further in future versions.

The testing includes the data input into the system and then will show if the result is what was intended. The testing will use the objectives in the Analysis section to see if the system does what is I set out to achieve.

TESTING – DATA ACQUISITION & CONSOLIDATION

1. Data Acquisition & Cleansing

Test	Objective	Process	Expected Output	Actual Output	Result
1	Sourcing website that has football player transfer market values.	Obtain transfer market website to obtain player transfer market values. Thorough analysis of html, data structure, URLs and output required. No simple 'download all data', therefore web scraping script will have to be developed	Csv output file of player ID's and their respective transfer market value	Csv output file of player ID's and their respective transfer market value	Pass
2	Sourcing website that has football player statistical attributes.	Obtain csvs from the fbref market website to obtain player attributes. Thorough analysis of html, data structure, URLs and output required. No simple 'download all data', therefore web scraping script will have to be made	Csv file of player ID's and their respective statistical attributes	Csv file of player ID's and their respective statistical attributes	Pass
3	Create dataset of players across multiple leagues.	In order to build a comprehensive system that analyses several data points, it is necessary to get player data across the top 5 European football leagues.	Data output file has data not just from one league but from the following leagues: (i) English Premier League (ii) Spanish La Liga (iii) German Bundesliga (iv) Italian Serie A (v) French Ligue 1	Consolidated csv data file across the top leagues	Pass
4	Create dataset of player attributes across multiple seasons. This will be a unique database from multiple sources, data points and years.	In order to build a system that accurately analyses several data points, it is necessary to get player data across multiple years and seasons.	Data output file has data not just from one season, but from the following seasons: (i) 2018/19 (ii) 2019/20 (iii) 2020/21	Three csv files: Consolidated_csv. Output is a unique database from multiple sources, data points and years.	Pass

TESTING - OUTPUT

Output of Tests: 1 - 4

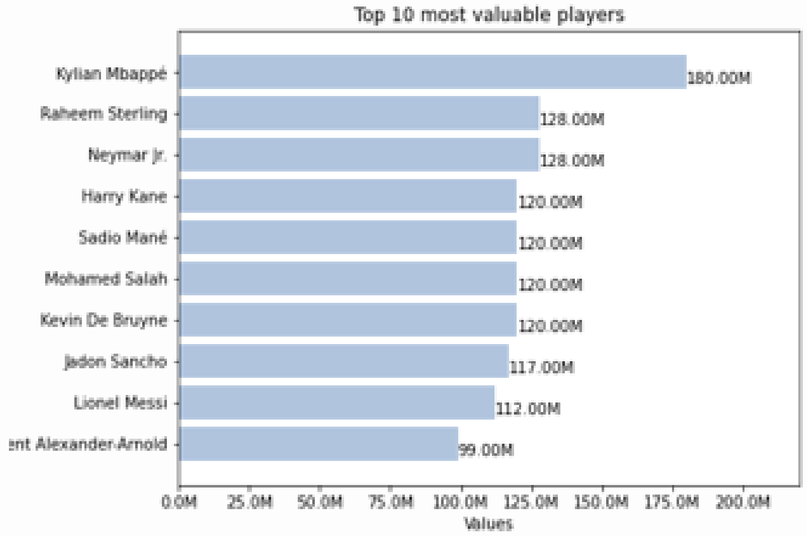
Jupyter server output showing the three data files. Each of these data files have thousands of records so it’s too large to display here, but the image from the jupyter server shows as a successful output. These datasets will be used as the basis for the data analysis and the machine learning.

TESTING – MOST VALUEABLE PLAYERS

2. Data Analytics

Test	Objective	Process	Expected Output	Actual Output	Result
5	List of most valuable players	Develop code to create a list of the most valuable players this season according to Transfermarkt. This will be used in the data analysis section	Jupyter server variables generates a data table sorted by the most valuable players	In the output, the list is correctly shown. The data file has been created and all the players are listed with values	Pass
6	Graphical output of top 10 most valuable players to be used in the final player report	Develop code to list the 10 players with the highest values according to Transfermarkt. This will be used in the data analysis section	Python code creates the table.	Graph created with player values of the top 10 most valuable players	Pass

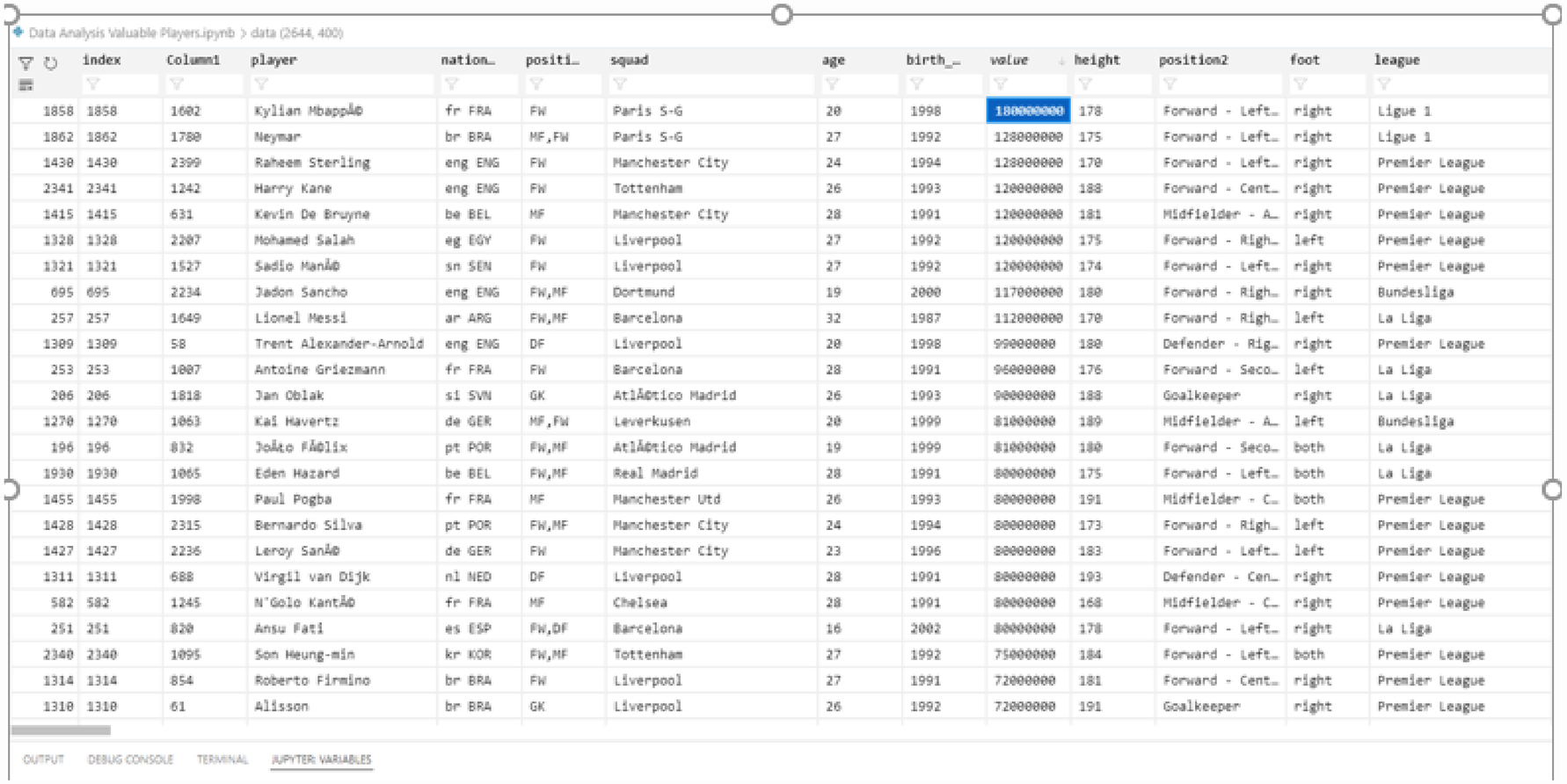
Test 6 – Output: Most Valuable Players



TESTING - OUTPUT

Output of Tests: 5 & 6

The jupyter server outputs the variables of consolidated player attributes and values. This is snapshot of the output after I ran the data scraping code. There are many records, but this snapshot from the juypter server shows the key columns for the valuable players. This data will be used in the Analytics report.



Data Analysis Valuable Players.ipynb > data (2644, 400)

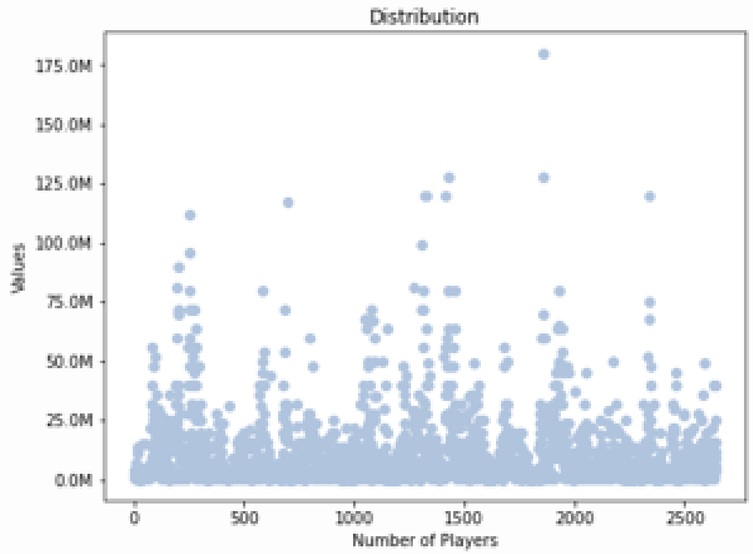
	index	Column1	player	nation…	positi…	squad	age	birth_…	value	height	position2	foot	league
1858	1858	1602	Kylian MbappÃ©	fr FRA	FW	Paris S-G	20	1998	180000000	178	Forward - Left…	right	Ligue 1
1862	1862	1780	Neymar	br BRA	MF,FW	Paris S-G	27	1992	128000000	175	Forward - Left…	right	Ligue 1
1430	1430	2399	Raheem Sterling	eng ENG	FW	Manchester City	24	1994	128000000	170	Forward - Left…	right	Premier League
2341	2341	1242	Harry Kane	eng ENG	FW	Tottenham	26	1993	120000000	188	Forward - Cent…	right	Premier League
1415	1415	631	Kevin De Bruyne	be BEL	MF	Manchester City	28	1991	120000000	181	Midfielder - A…	right	Premier League
1328	1328	2207	Mohamed Salah	eg EGY	FW	Liverpool	27	1992	120000000	175	Forward - Righ…	left	Premier League
1321	1321	1527	Sadio ManÃ©	sn SEN	FW	Liverpool	27	1992	120000000	174	Forward - Left…	right	Premier League
695	695	2234	Jadon Sancho	eng ENG	FW,MF	Dortmund	19	2000	117000000	180	Forward - Righ…	right	Bundesliga
257	257	1649	Lionel Messi	ar ARG	FW,MF	Barcelona	32	1987	112000000	170	Forward - Righ…	left	La Liga
1309	1309	58	Trent Alexander-Arnold	eng ENG	DF	Liverpool	20	1998	99000000	180	Defender - Rig…	right	Premier League
253	253	1007	Antoine Griezmann	fr FRA	FW	Barcelona	28	1991	96000000	176	Forward - Seco…	left	La Liga
206	206	1818	Jan Oblak	si SVN	GK	AtlÃ©tico Madrid	26	1993	90000000	188	Goalkeeper	right	La Liga
1270	1270	1063	Kai Havertz	de GER	MF,FW	Leverkusen	20	1999	81000000	189	Midfielder - A…	left	Bundesliga
196	196	832	JoÃ£o FÃ©lix	pt POR	FW,MF	AtlÃ©tico Madrid	19	1999	81000000	180	Forward - Seco…	both	La Liga
1930	1930	1065	Eden Hazard	be BEL	FW,MF	Real Madrid	28	1991	80000000	175	Forward - Left…	both	La Liga
1455	1455	1998	Paul Pogba	fr FRA	MF	Manchester Utd	26	1993	80000000	191	Midfielder - C…	both	Premier League
1428	1428	2315	Bernardo Silva	pt POR	FW,MF	Manchester City	24	1994	80000000	173	Forward - Righ…	left	Premier League
1427	1427	2236	Leroy SanÃ©	de GER	FW	Manchester City	23	1996	80000000	183	Forward - Left…	left	Premier League
1311	1311	688	Virgil van Dijk	nl NED	DF	Liverpool	28	1991	80000000	193	Defender - Cen…	right	Premier League
582	582	1245	N'Golo KantÃ©	fr FRA	MF	Chelsea	28	1991	80000000	168	Midfielder - C…	right	Premier League
251	251	820	Ansu Fati	es ESP	FW,DF	Barcelona	16	2002	80000000	178	Forward - Left…	right	La Liga
2340	2340	1095	Son Heung-min	kr KOR	FW,MF	Tottenham	27	1992	75000000	184	Forward - Left…	both	Premier League
1314	1314	854	Roberto Firmino	br BRA	FW	Liverpool	27	1991	72000000	181	Forward - Cent…	right	Premier League
1310	1310	61	Alisson	br BRA	GK	Liverpool	26	1992	72000000	191	Goalkeeper	right	Premier League

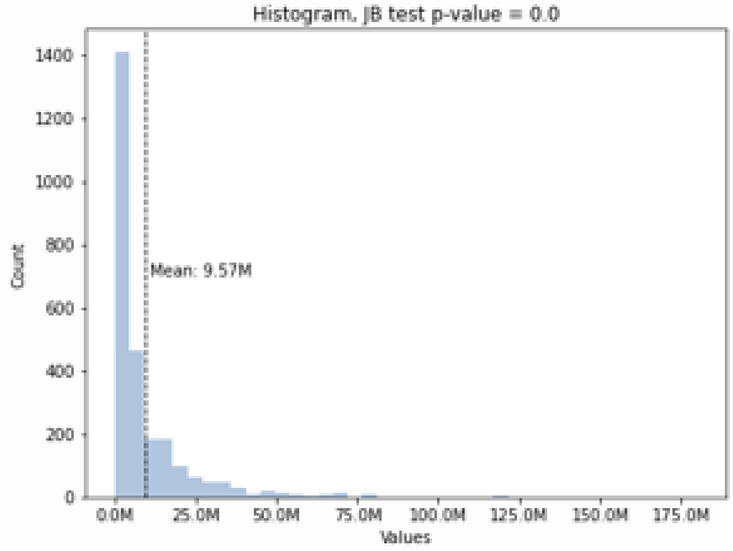
OUTPUT DEBUG CONSOLE TERMINAL JUPYTER VARIABLES

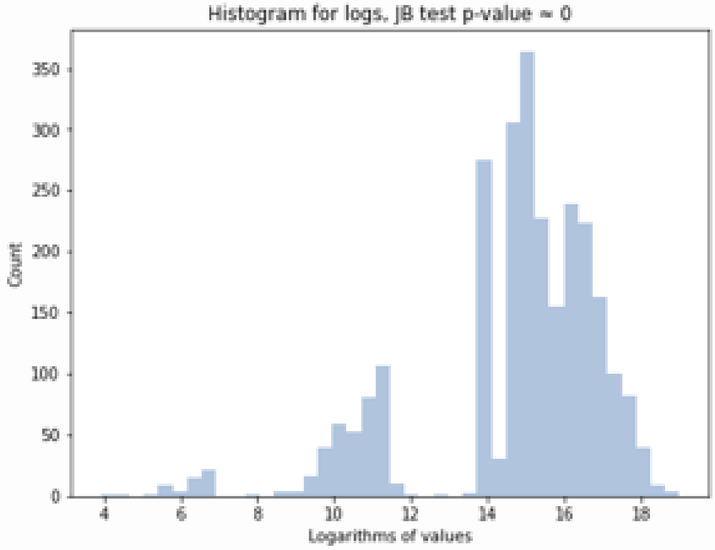
TESTING – DISTRIBUTION & LOG OF PLAYER VALUES

Test	Objective	Process	Expected Output	Actual Output	Result
7	Graph showing the distribution of the value of all players vs the number of players	Develop code to show the distribution of the values of all players. This will be used in the player report.	Distribution of values to frequency table	Graph showing distribution of values generated	Pass
8	Histograms of the value of all players and logarithms of the values vs their transfer values.	Develop code to show two histograms of the values of all players and the logarithms of values vs count. This will be used in the player report.	Histograms of value/ log of values vs number of players.	Histograms of value/ log of values vs number of players.	Pass

Test 7 & 8 – Output: Distribution & Histogram of all Players





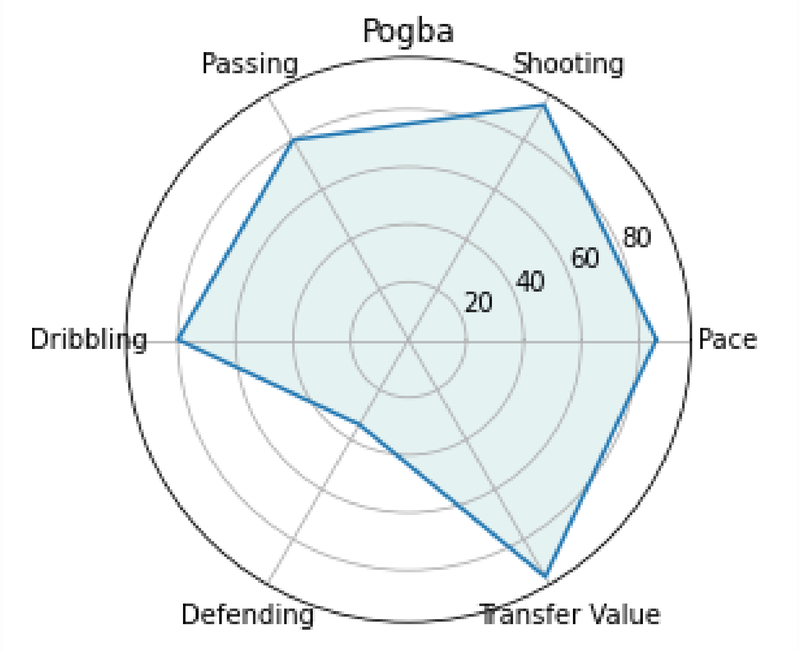


TESTING – SINGLE PLAYER ANALYSIS

2. Data Analytics – Player Comparative Analysis – Single Player

Test	Objective	Process	Expected Output	Actual Output	Result
9	Show graphically the abilities of a player including their transfer value to give the user a way to visualise the data.	Develop code to create a graphical list of a player's abilities and the player's transfer value. These are taken from the consolidated database. Then they are used to create a spider pie chart. This will be used in the data analysis report.	Graphical spider pie chart	Graphical spider pie chart	Pass

Test 9 – Output: Player Comparative Analysis – Single Player

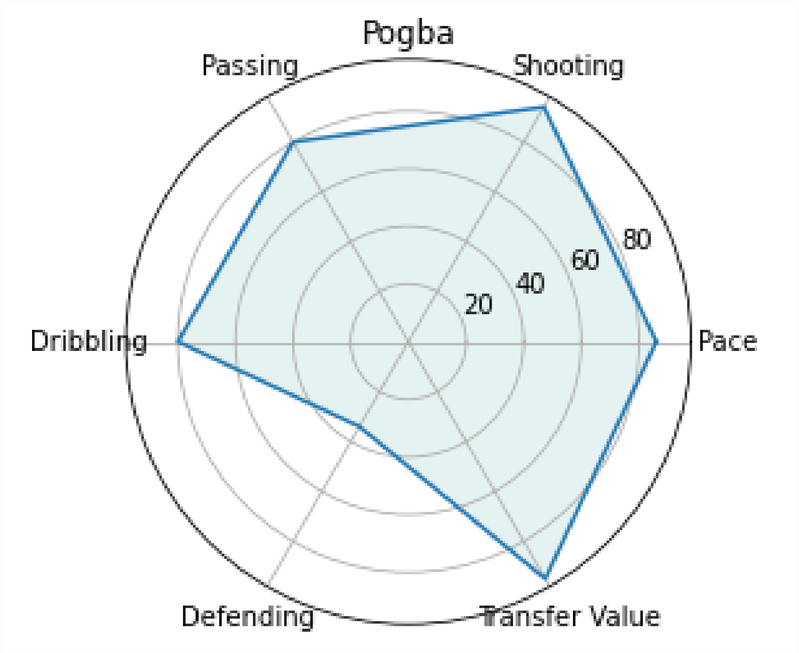


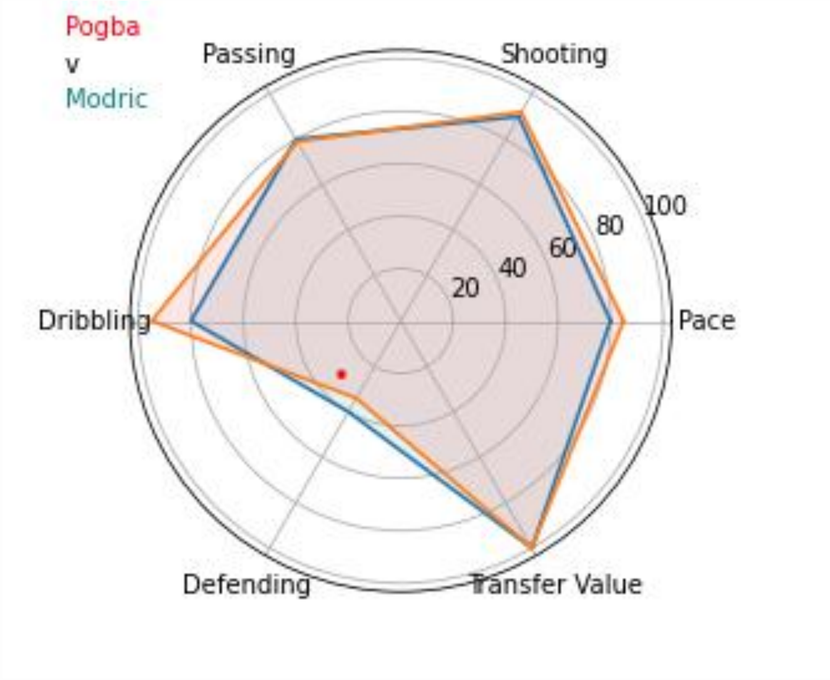
TESTING – COMPARITIVE PLAYER ANALYSIS

2. Data Analytics – Player Comparative Analysis

Test	Objective	Process	Expected Output	Actual Output	Result
10	Compare graphically two players. This will be very useful for the user to compare two player's abilities and transfer values.	Develop code to compare graphically two players. The spider pie will overlay one player's abilities and transfer value over the other. This will be used in the data analysis report.	Graphical spider pie chart showing the two player's abilities overlayed on one another.	Graphical spider pie chart showing the two player's abilities overlayed on one another.	Pass

Test 9 – Output: Player Comparative Analysis – Comparison of Player



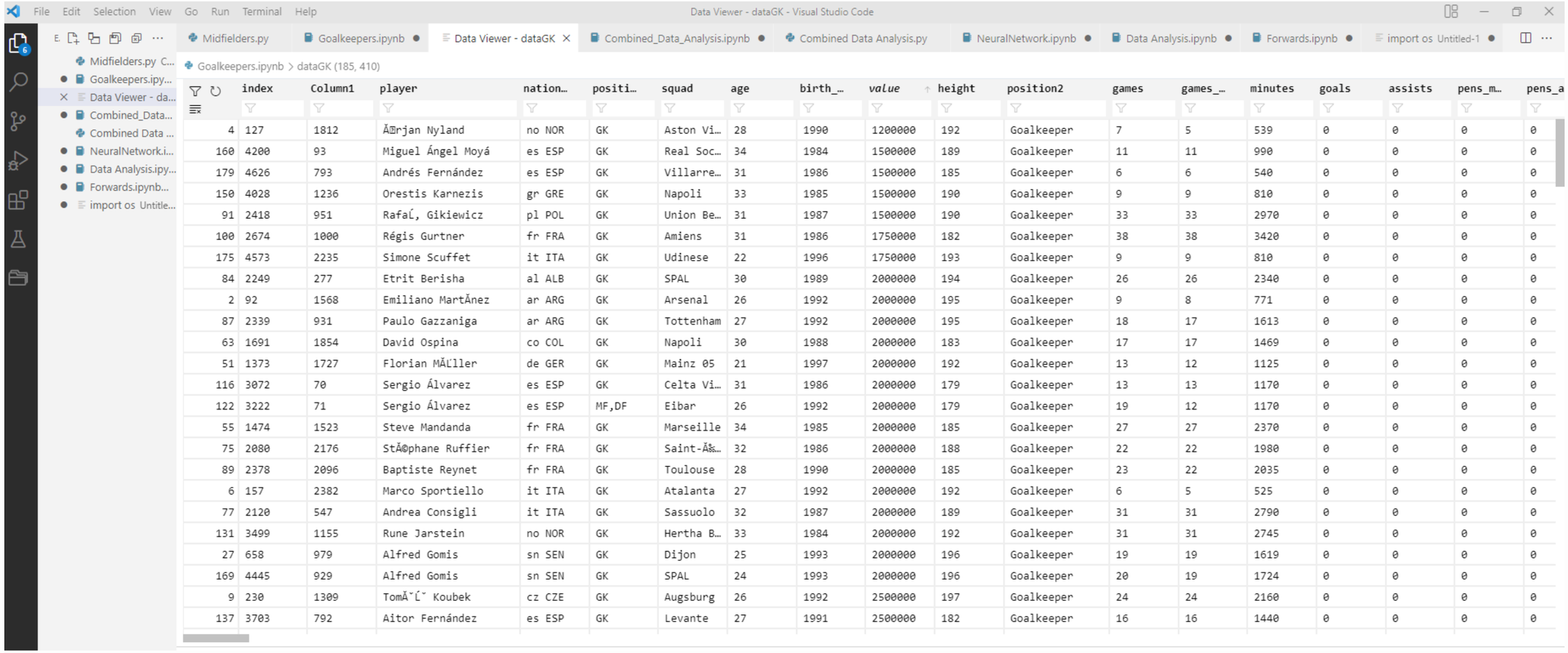


TESTING – LINEAR REGRESSION

3. Machine Learning – Dataset Goalkeepers

Test	Objective	Process	Expected Output	Actual Output	Result
11	The first part of the machine learning is to create a set of players for all players in that position. The first analysis is for all goalkeepers.	Develop code to import the consolidated data for each of the three seasons. Select goalkeepers from that dataset. Remove outliers in dataset.	Dataset for goalkeepers	Jupyter output below is the snapshot of the dataset for goalkeepers	Pass

Test 11 – Output: Dataset – Goalkeepers



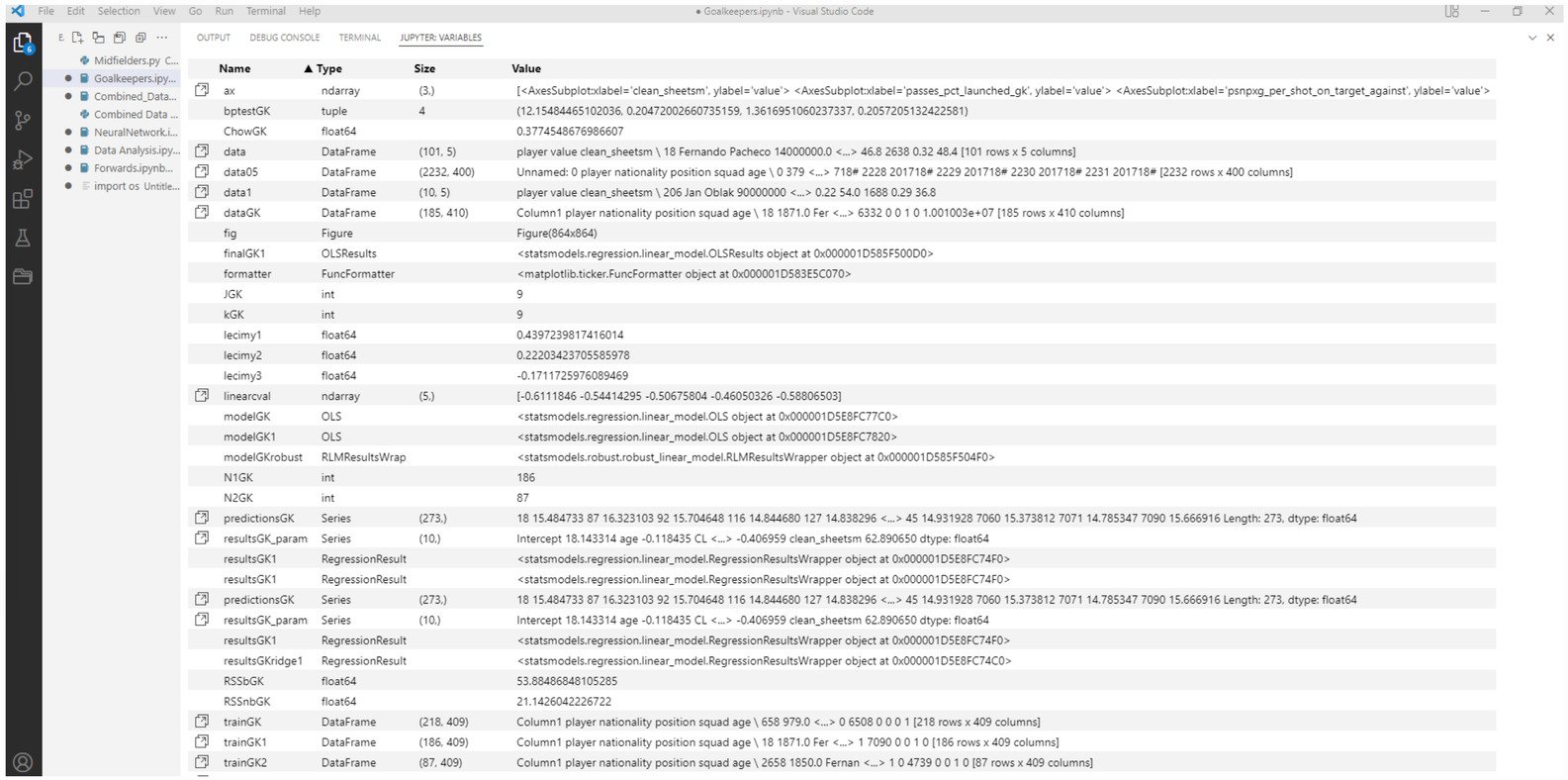
	index	Column1	player	nation…	positi…	squad	age	birth_…	value	height	position2	games	games_…	minutes	goals	assists	pens_m…	pens_a
4	127	1812	Ã Rjan Nyland	no NOR	GK	Aston Vi…	28	1990	1200000	192	Goalkeeper	7	5	539	0	0	0	0
160	4200	93	Miguel Ángel Moyá	es ESP	GK	Real Soc…	34	1984	1500000	189	Goalkeeper	11	11	990	0	0	0	0
179	4626	793	Andrés Fernández	es ESP	GK	Villarre…	31	1986	1500000	185	Goalkeeper	6	6	540	0	0	0	0
150	4028	1236	Orestis Karnezis	gr GRE	GK	Napoli	33	1985	1500000	190	Goalkeeper	9	9	810	0	0	0	0
91	2418	951	Rafaĺ, Gikiewicz	pl POL	GK	Union Be…	31	1987	1500000	190	Goalkeeper	33	33	2970	0	0	0	0
100	2674	1000	Régis Gurtner	fr FRA	GK	Amiens	31	1986	1750000	182	Goalkeeper	38	38	3420	0	0	0	0
175	4573	2235	Simone Scuffet	it ITA	GK	Udinese	22	1996	1750000	193	Goalkeeper	9	9	810	0	0	0	0
84	2249	277	Etrit Berisha	al ALB	GK	SPAL	30	1989	2000000	194	Goalkeeper	26	26	2340	0	0	0	0
2	92	1568	Emiliano MartÃnez	ar ARG	GK	Arsenal	26	1992	2000000	195	Goalkeeper	9	8	771	0	0	0	0
87	2339	931	Paulo Gazzaniga	ar ARG	GK	Tottenham	27	1992	2000000	195	Goalkeeper	18	17	1613	0	0	0	0
63	1691	1854	David Ospina	co COL	GK	Napoli	30	1988	2000000	183	Goalkeeper	17	17	1469	0	0	0	0
51	1373	1727	Florian MÄĽller	de GER	GK	Mainz 05	21	1997	2000000	192	Goalkeeper	13	12	1125	0	0	0	0
116	3072	70	Sergio Álvarez	es ESP	GK	Celta Vi…	31	1986	2000000	179	Goalkeeper	13	13	1170	0	0	0	0
122	3222	71	Sergio Álvarez	es ESP	MF,DF	Eibar	26	1992	2000000	179	Goalkeeper	19	12	1170	0	0	0	0
55	1474	1523	Steve Mandanda	fr FRA	GK	Marseille	34	1985	2000000	185	Goalkeeper	27	27	2370	0	0	0	0
75	2080	2176	StÃ©phane Ruffier	fr FRA	GK	Saint-Ã‰…	32	1986	2000000	188	Goalkeeper	22	22	1980	0	0	0	0
89	2378	2096	Baptiste Reynet	fr FRA	GK	Toulouse	28	1990	2000000	185	Goalkeeper	23	22	2035	0	0	0	0
6	157	2382	Marco Sportiello	it ITA	GK	Atalanta	27	1992	2000000	192	Goalkeeper	6	5	525	0	0	0	0
77	2120	547	Andrea Consigli	it ITA	GK	Sassuolo	32	1987	2000000	189	Goalkeeper	31	31	2790	0	0	0	0
131	3499	1155	Rune Jarstein	no NOR	GK	Hertha B…	33	1984	2000000	192	Goalkeeper	31	31	2745	0	0	0	0
27	658	979	Alfred Gomis	sn SEN	GK	Dijon	25	1993	2000000	196	Goalkeeper	19	19	1619	0	0	0	0
169	4445	929	Alfred Gomis	sn SEN	GK	SPAL	24	1993	2000000	196	Goalkeeper	20	19	1724	0	0	0	0
9	230	1309	TomÃ Ĺ Koubek	cz CZE	GK	Augsburg	26	1992	2500000	197	Goalkeeper	24	24	2160	0	0	0	0
137	3703	792	Aitor Fernández	es ESP	GK	Levante	27	1991	2500000	182	Goalkeeper	16	16	1440	0	0	0	0

TESTING – LINEAR REGRESSION

3. Machine Learning - Linear Regression - Goalkeepers

Test	Objective	Process	Expected Output	Actual Output	Result
12	To create linear regression on a goalkeeper dataset using variables that are important in determining a goalkeeper's value.	Linear regression - comparing value of the goalkeeper to the variables that influence a goalkeeper's value: age, league, clean sheets, wins and minutes played variables . Create test and training set and run linear regression model	DataGK TrainGK PredictionsGK	DataGK TrainGK PredictionsGK	Pass

Test 12 – Output: Dataset – Goalkeepers



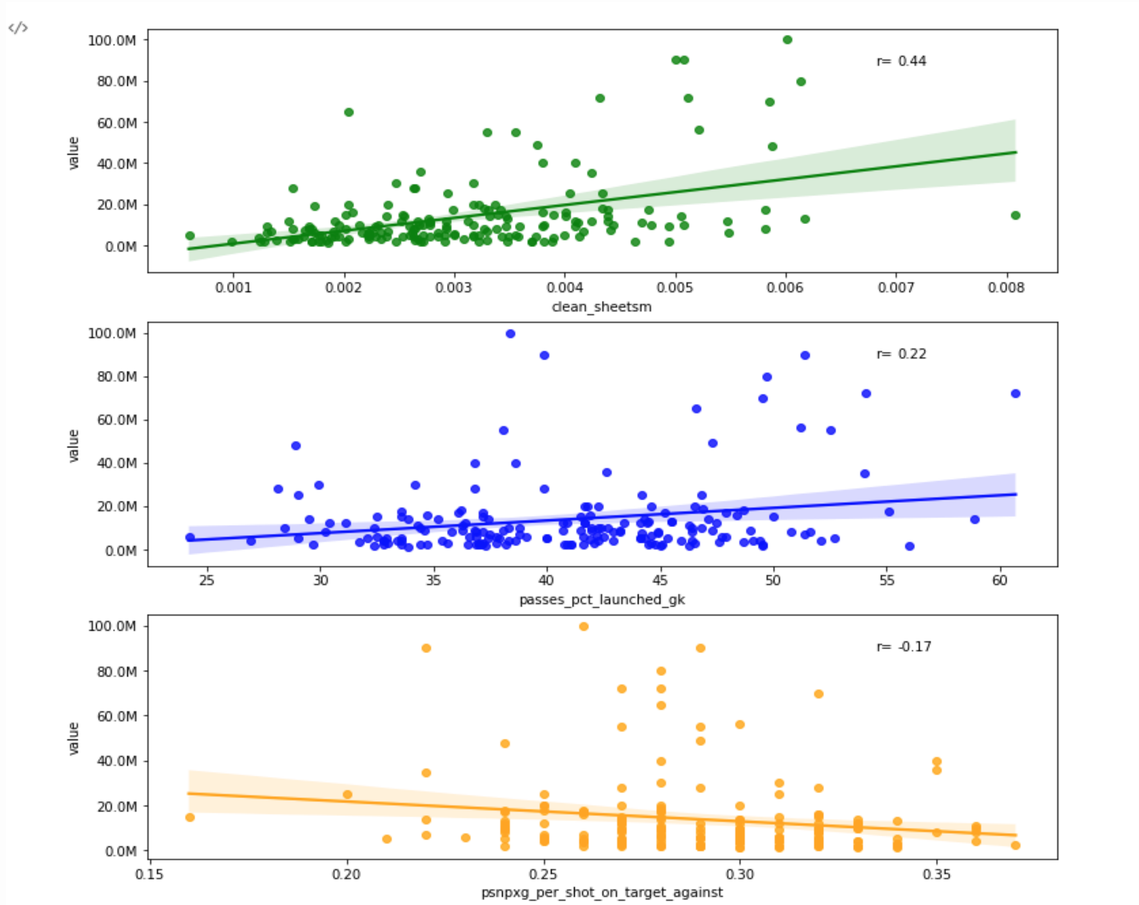
Name	Type	Size	Value
ax	ndarray	(3,)	[<AxesSubplot:xlabel='clean_sheetsm', ylabel='value'> <AxesSubplot:xlabel='passes_pct_launched_gk', ylabel='value'> <AxesSubplot:xlabel='psnpxg_per_shot_on_target_against', ylabel='value'>
bptestGK	tuple	4	(12.15484465102036, 0.20472002660735159, 1.3616951060237337, 0.2057205132422581)
ChowGK	float64		0.3774548676986607
data	DataFrame	(101, 5)	player value clean_sheetsm \ 18 Fernando Pacheco 14000000.0 <...> 46.8 2638 0.32 48.4 [101 rows x 5 columns]
data05	DataFrame	(2232, 400)	Unnamed: 0 player nationality position squad age \ 0 379 <...> 718# 2228 201718# 2229 201718# 2230 201718# 2231 201718# [2232 rows x 400 columns]
data1	DataFrame	(10, 5)	player value clean_sheetsm \ 206 Jan Oblak 90000000 <...> 0.22 54.0 1688 0.29 36.8
dataGK	DataFrame	(185, 410)	Column1 player nationality position squad age \ 18 1871.0 Fer <...> 6332 0 0 1 0 1.001003e+07 [185 rows x 410 columns]
fig	Figure		Figure(864x864)
finalGK1	OLSResults		<statsmodels.regression.linear_model.OLSResults object at 0x000001D585F500D0>
formatter	FuncFormatter		<matplotlib.ticker.FuncFormatter object at 0x000001D583E5C070>
JGK	int		9
kGK	int		9
lecimy1	float64		0.4397239817416014
lecimy2	float64		0.22203423705585978
lecimy3	float64		-0.1711725976089469
linearcval	ndarray	(5,)	[-0.6111846 -0.54414295 -0.50675804 -0.46050326 -0.58806503]
modelGK	OLS		<statsmodels.regression.linear_model.OLS object at 0x000001D5E8FC77C0>
modelGK1	OLS		<statsmodels.regression.linear_model.OLS object at 0x000001D5E8FC7820>
modelGKrobust	RLMResultsWrap		<statsmodels.robust.robust_linear_model.RLMResultsWrapper object at 0x000001D585F504F0>
N1GK	int		186
N2GK	int		87
predictionsGK	Series	(273,)	18 15.484733 87 16.323103 92 15.704648 116 14.844680 127 14.838296 <...> 45 14.931928 7060 15.373812 7071 14.785347 7090 15.666916 Length: 273, dtype: float64
resultsGK_param	Series	(10,)	Intercept 18.143314 age -0.118435 CL <...> -0.406959 clean_sheetsm 62.890650 dtype: float64
resultsGK1	RegressionResult		<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x000001D5E8FC74F0>
resultsGK1	RegressionResult		<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x000001D5E8FC74F0>
predictionsGK	Series	(273,)	18 15.484733 87 16.323103 92 15.704648 116 14.844680 127 14.838296 <...> 45 14.931928 7060 15.373812 7071 14.785347 7090 15.666916 Length: 273, dtype: float64
resultsGK_param	Series	(10,)	Intercept 18.143314 age -0.118435 CL <...> -0.406959 clean_sheetsm 62.890650 dtype: float64
resultsGK1	RegressionResult		<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x000001D5E8FC74F0>
resultsGKridge1	RegressionResult		<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x000001D5E8FC74C0>
RSSbGK	float64		53.88486848105285
RSSnbGK	float64		21.1426042226722
trainGK	DataFrame	(218, 409)	Column1 player nationality position squad age \ 658 979.0 <...> 0 6508 0 0 0 1 [218 rows x 409 columns]
trainGK1	DataFrame	(186, 409)	Column1 player nationality position squad age \ 18 1871.0 Fer <...> 1 7090 0 0 1 0 [186 rows x 409 columns]
trainGK2	DataFrame	(87, 409)	Column1 player nationality position squad age \ 2658 1850.0 Fernan <...> 1 0 4739 0 0 1 0 [87 rows x 409 columns]

TESTING – LINEAR REGRESSION

3. Machine Learning - Linear Regression – Goalkeepers - Output

Test	Objective	Process	Expected Output	Actual Output	Result
13	To create a graphical output of the goalkeepers' linear regression of values vs the three key variables that influence a goalkeeper's value.	Create a data plot using python stats to show key variables for the value: clean sheet, passes and shots against	Linear regression plot of value to variables	Linear regression for each of the three key goalkeeper variables.	Pass

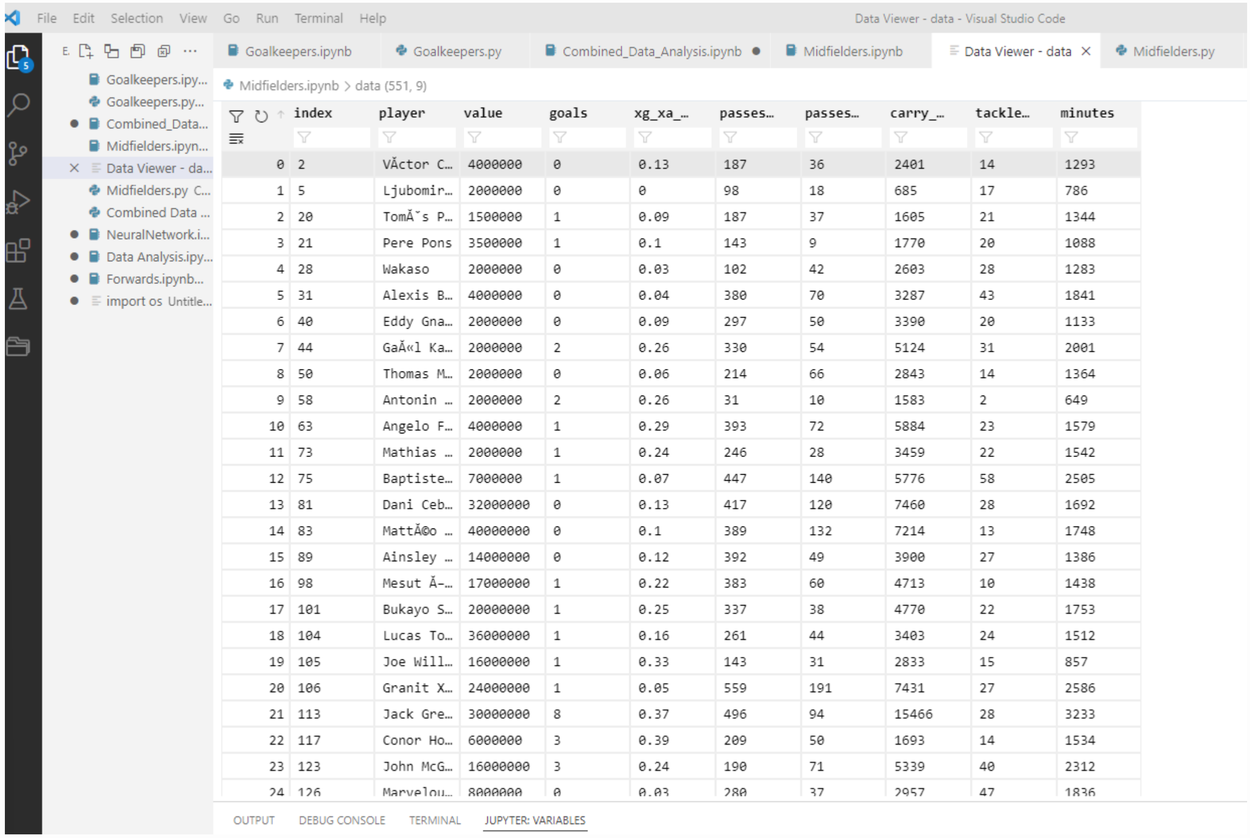
Test 13 – Output: Dataset – Goalkeepers



TESTING – LINEAR REGRESSION - MIDFIELDERS

3. Machine Learning - Linear Regression – Midfielders - Output

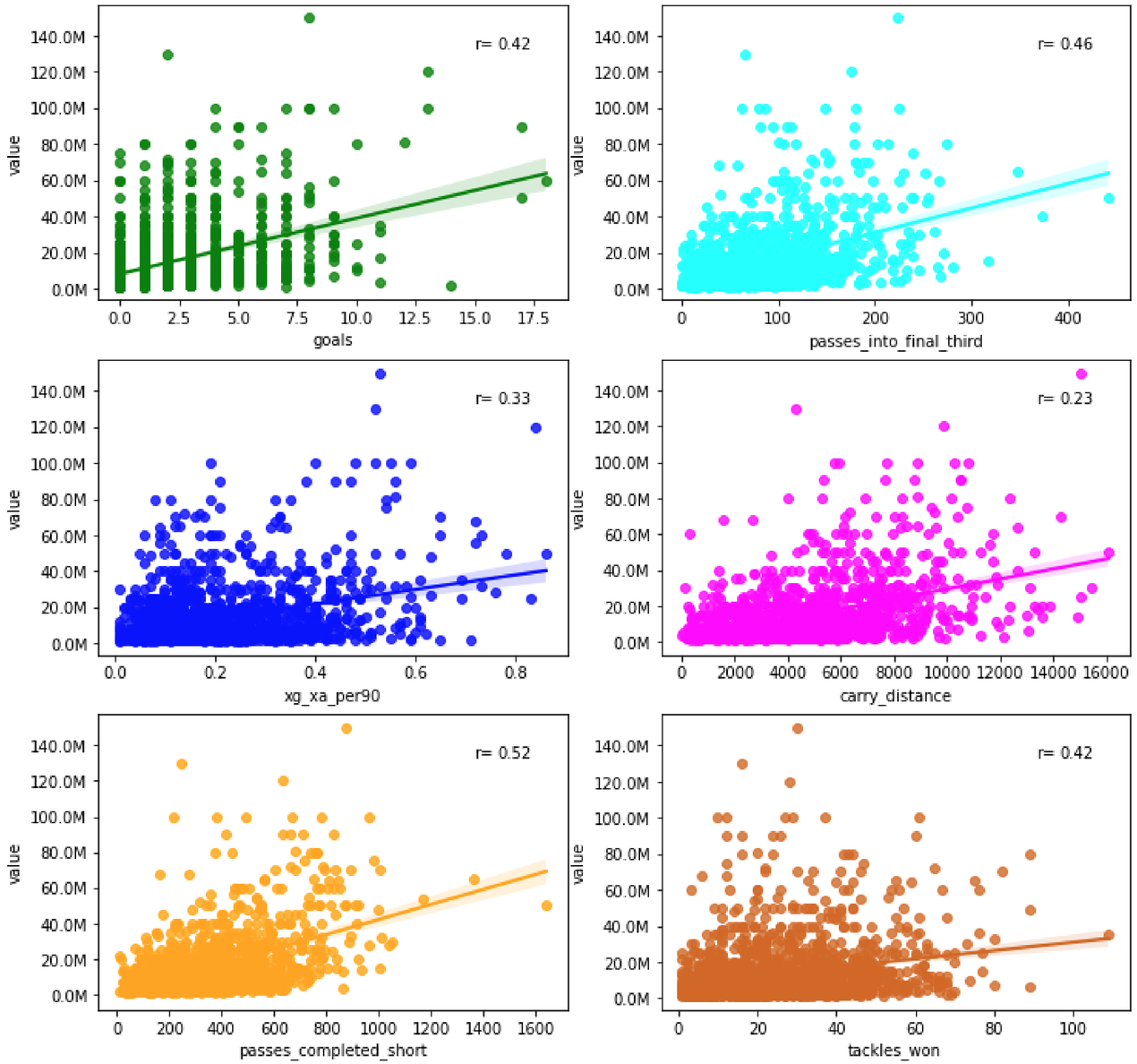
Test	Objective	Process	Expected Output	Actual Output	Result
14	The first part of the machine learning is to create a set of players for all players in that position. The next analysis is for all midfielders.	Create data plot of key variables for the value: clean sheet, passes and shots against#dataMID=dataMID[['goals','xg_xa_per90','passes_completed_short','passes_into_final_third','carry_distance','tackles_won']	Linear regression plot of value to variables	Linear regression for each of the key midfielder variables.	Pass



	index	player	value	goals	xg_xa_...	passes...	passes...	carry_...	tackle...	minutes
0	2	VÃctor C...	4000000	0	0.13	187	36	2401	14	1293
1	5	Ljubomir...	2000000	0	0	98	18	685	17	786
2	20	TomÃ¯s P...	1500000	1	0.09	187	37	1605	21	1344
3	21	Pere Pons	3500000	1	0.1	143	9	1770	20	1088
4	28	Wakaso	2000000	0	0.03	102	42	2603	28	1283
5	31	Alexis B...	4000000	0	0.04	380	70	3287	43	1841
6	40	Eddy Gna...	2000000	0	0.09	297	50	3390	20	1133
7	44	GaÃ«l Ka...	2000000	2	0.26	330	54	5124	31	2001
8	50	Thomas M...	2000000	0	0.06	214	66	2843	14	1364
9	58	Antonin ...	2000000	2	0.26	31	10	1583	2	649
10	63	Angelo F...	4000000	1	0.29	393	72	5884	23	1579
11	73	Mathias ...	2000000	1	0.24	246	28	3459	22	1542
12	75	Baptiste...	7000000	1	0.07	447	140	5776	58	2505
13	81	Dani Ceb...	32000000	0	0.13	417	120	7460	28	1692
14	83	MattÃ©o ...	40000000	0	0.1	389	132	7214	13	1748
15	89	Ainsley ...	14000000	0	0.12	392	49	3900	27	1386
16	98	Mesut Ã-...	17000000	1	0.22	383	60	4713	10	1438
17	101	Bukayo S...	20000000	1	0.25	337	38	4770	22	1753
18	104	Lucas To...	36000000	1	0.16	261	44	3403	24	1512
19	105	Joe Will...	16000000	1	0.33	143	31	2833	15	857
20	106	Granit X...	24000000	1	0.05	559	191	7431	27	2586
21	113	Jack Gre...	30000000	8	0.37	496	94	15466	28	3233
22	117	Conor Ho...	6000000	3	0.39	209	50	1693	14	1534
23	123	John McG...	16000000	3	0.24	190	71	5339	40	2312
24	126	Marvelou...	8000000	0	0.03	280	37	2957	47	1836

TESTING – LINEAR REGRESSION - MIDFIELDERS

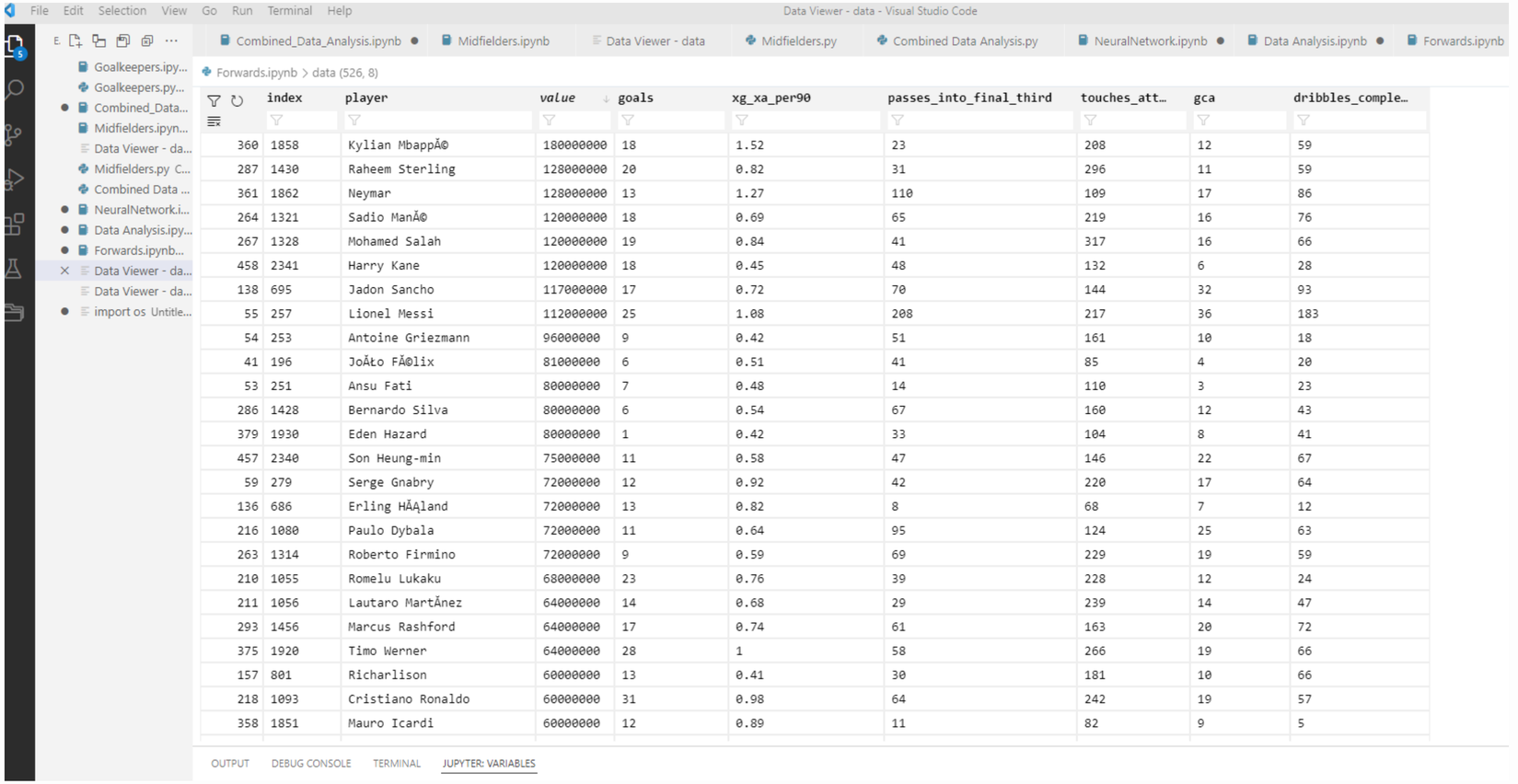
Test 14 – Output: Dataset – Midfielders



TESTING – LINEAR REGRESSION - FORWARDS

3. Machine Learning - Linear Regression – Forwards - Output

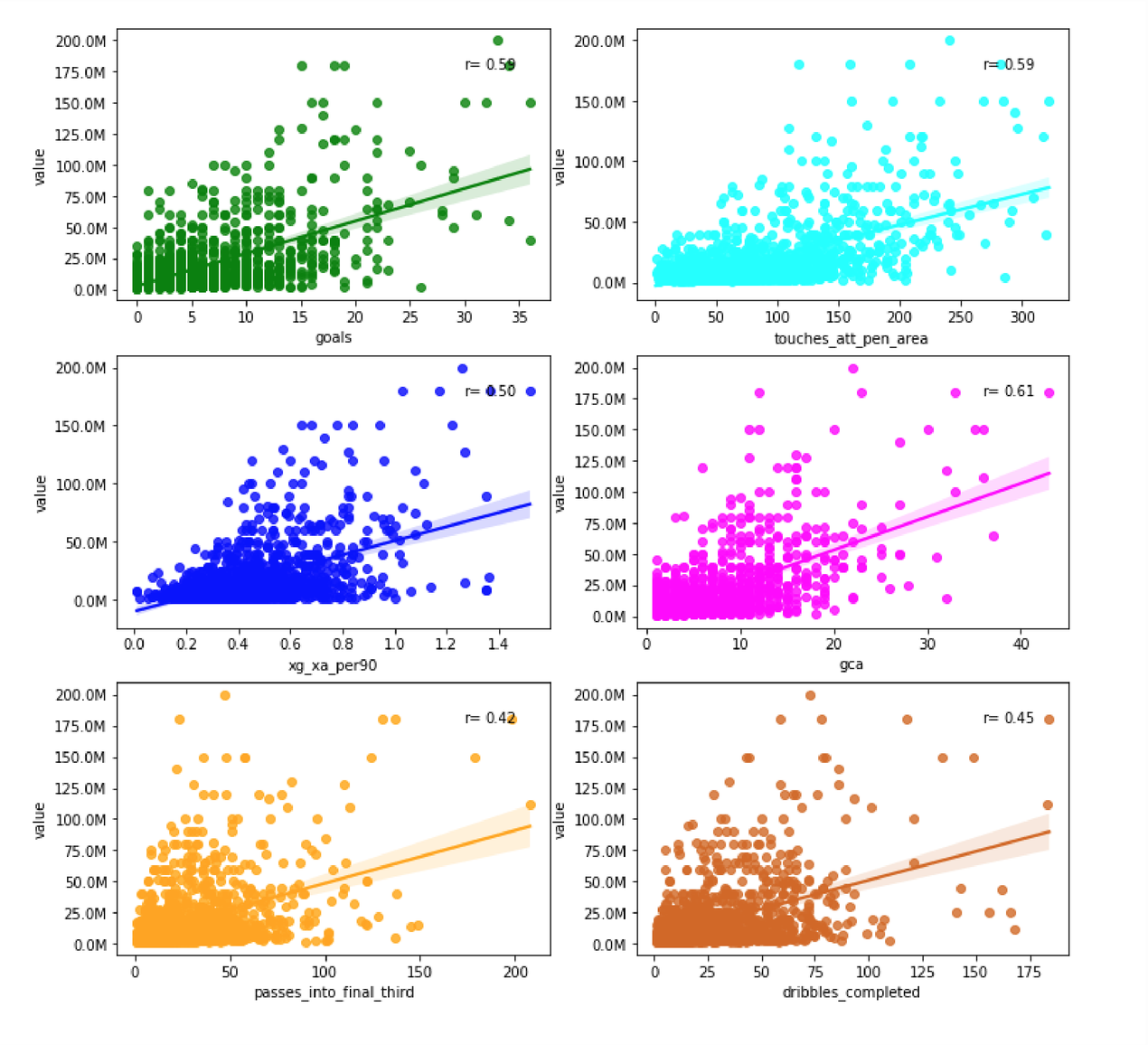
Test	Objective	Process	Expected Output	Actual Output	Result
15	To create the output of the forwards linear regression for the key variables that influence a forwards value.	Create linear regression plot of key variables for forwards: age+CL+goals+gca' '+Pts+xG+xGA+dribbles_completed' '+xg_xa_per90+touches_att_pen_area+' '+passes_into_final_third+' '+isPremierLeague+isLigue1	Linear regression plot of values of forwards to the distribution of the values	Linear regression plot of values of forwards to the distribution of the values	Pass



	index	player	value	goals	xg_xa_per90	passes_into_final_third	touches_att…	gca	dribbles_comple…
360	1858	Kylian MbappÃ©	180000000	18	1.52	23	208	12	59
287	1430	Raheem Sterling	128000000	20	0.82	31	296	11	59
361	1862	Neymar	128000000	13	1.27	110	109	17	86
264	1321	Sadio ManÃ©	120000000	18	0.69	65	219	16	76
267	1328	Mohamed Salah	120000000	19	0.84	41	317	16	66
458	2341	Harry Kane	120000000	18	0.45	48	132	6	28
138	695	Jadon Sancho	117000000	17	0.72	70	144	32	93
55	257	Lionel Messi	112000000	25	1.08	208	217	36	183
54	253	Antoine Griezmann	96000000	9	0.42	51	161	10	18
41	196	JoÃ£o FÃ©lix	81000000	6	0.51	41	85	4	20
53	251	Ansu Fati	80000000	7	0.48	14	110	3	23
286	1428	Bernardo Silva	80000000	6	0.54	67	160	12	43
379	1930	Eden Hazard	80000000	1	0.42	33	104	8	41
457	2340	Son Heung-min	75000000	11	0.58	47	146	22	67
59	279	Serge Gnabry	72000000	12	0.92	42	220	17	64
136	686	Erling HÃ¥land	72000000	13	0.82	8	68	7	12
216	1080	Paulo Dybala	72000000	11	0.64	95	124	25	63
263	1314	Roberto Firmino	72000000	9	0.59	69	229	19	59
210	1055	Romelu Lukaku	68000000	23	0.76	39	228	12	24
211	1056	Lautaro MartÃnez	64000000	14	0.68	29	239	14	47
293	1456	Marcus Rashford	64000000	17	0.74	61	163	20	72
375	1920	Timo Werner	64000000	28	1	58	266	19	66
157	801	Richarlison	60000000	13	0.41	30	181	10	66
218	1093	Cristiano Ronaldo	60000000	31	0.98	64	242	19	57
358	1851	Mauro Icardi	60000000	12	0.89	11	82	9	5

TESTING – LINEAR REGRESSION - FORWARDS

Test 14 – Output: Dataset – Forwards



TESTING – PREDICT PLAYER VALUE

4. Machine Learning – Predicted Player Value

This is one of the most important tests in the project. It is the output of the machine learning algorithm. In the following test, I have run a series of tests showing the output from the neural network.

Test	Objective	Process	Expected Output	Actual Output	Result
16	This is one of the most interesting tests in the project. It is the output of the regression machine learning algorithm which gives the 'predicted' value of a player which can be compared to their current transfer market value to ascertain if over or under valued.	Using Jupyter server (i) Import my goalkeeper.py code (ii) Import my midfielder.py code (iii) Import my forward.py code Apply weights to each of the three seasons data: (i) 1 (ii) 0.8 (iii) 0.7 Append to dataset. Select a player and test. For this test I have selected a Man Utd player: "Paul Pogba".	The player value for Paul Pogba according to the Transfermarkt website is 100,000,000 The 'predictOLS' value is 98.129,368	The 'predictOLS' value of the player is: 98.129,368 predsOLS 98129368.	Pass

Test 16 – Output: Predicted Player Value

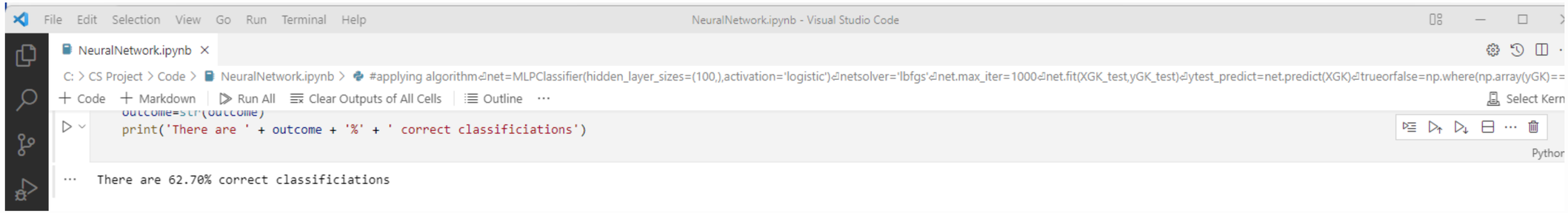
player	nation…	positi…	squad	age	birth_…	value	predsOLS	positi…	games	games_star…	minutes	goals	assists
pogba									3.				
Paul Pogba	fr FRA	MF	Manchester Utd	25	1993	100000000	98129368.	Midfield…	35	34	3006	13	9

TESTING – NEURAL NETWORK PREDICTIONS

4. Neural Network – MLPClassifer Predictions

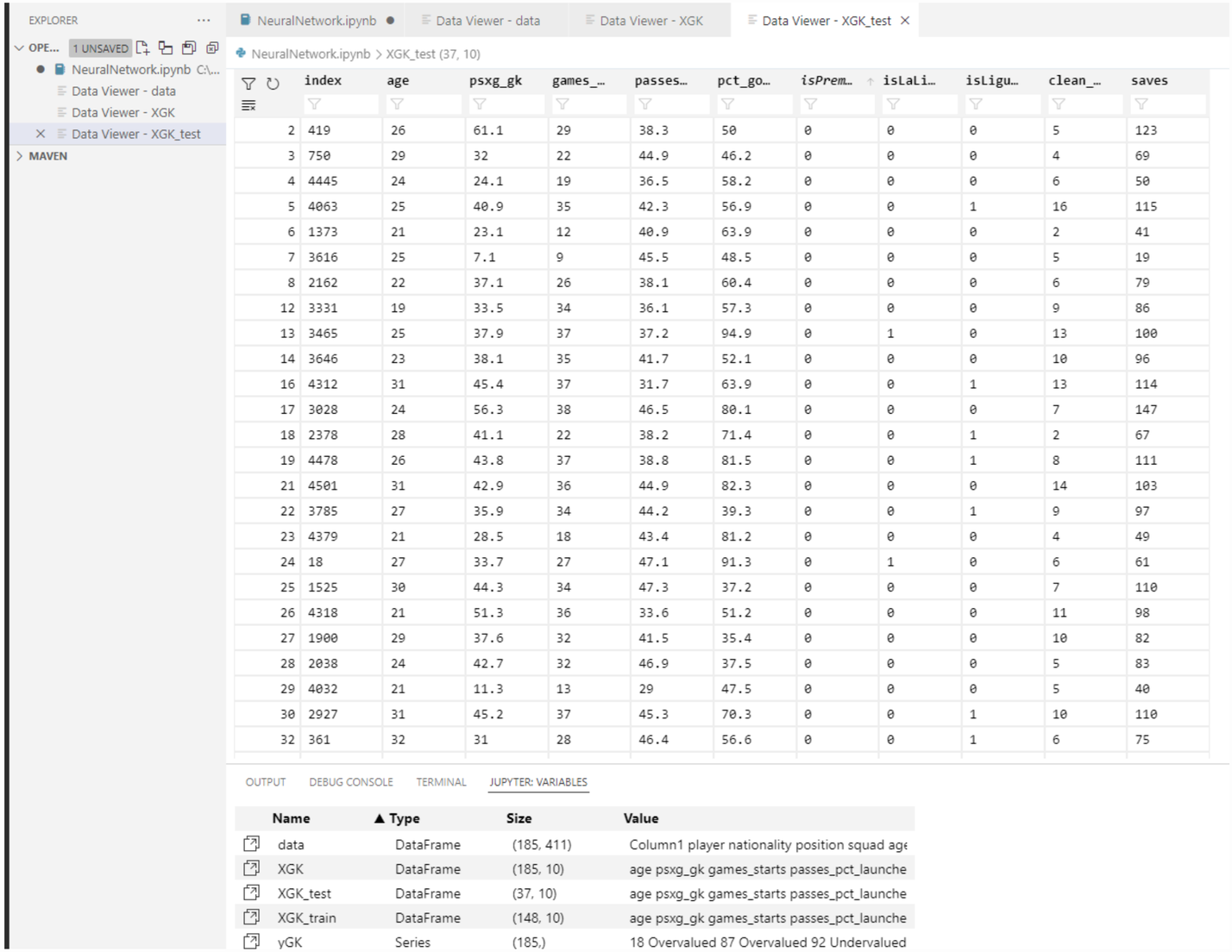
Test	Objective	Process	Expected Output	Actual Output	Result
17	This test is assessing how good the neural network is at predicting the value.	I will take my goalkeeper.py code, define the key variables for a goalkeeper and with training data use the MLPClassifer to show the % of output that is the correctly classified. Import my goalkeepers.py code. yGK = data['Over/undervalued'] XGK_train, XGK_test, yGK,test_size=0.2) Apply MLPClassifer Apply netsolver	Expected output will be a % of correct classifications	62.70 % of correct predictions – see the output of the neural network	Pass

Test 17 – Output: Neural Network Predictions



TESTING – NEURAL NETWORK PREDICTIONS

Test 17 – Output: Neural Network Predictions

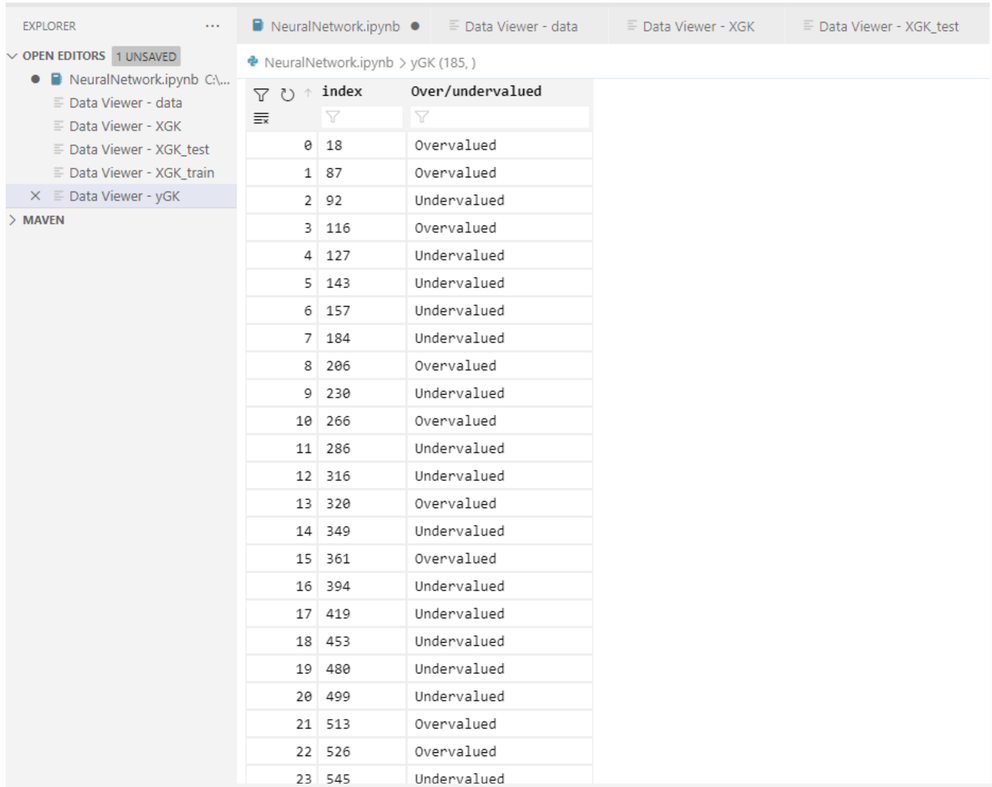


	index	age	psxg_gk	games_...	passes...	pct_go...	isPrem...	isLaLi...	isLigu...	clean_...	saves
2	419	26	61.1	29	38.3	50	0	0	0	5	123
3	750	29	32	22	44.9	46.2	0	0	0	4	69
4	4445	24	24.1	19	36.5	58.2	0	0	0	6	50
5	4063	25	40.9	35	42.3	56.9	0	0	1	16	115
6	1373	21	23.1	12	40.9	63.9	0	0	0	2	41
7	3616	25	7.1	9	45.5	48.5	0	0	0	5	19
8	2162	22	37.1	26	38.1	60.4	0	0	0	6	79
12	3331	19	33.5	34	36.1	57.3	0	0	0	9	86
13	3465	25	37.9	37	37.2	94.9	0	1	0	13	100
14	3646	23	38.1	35	41.7	52.1	0	0	0	10	96
16	4312	31	45.4	37	31.7	63.9	0	0	1	13	114
17	3028	24	56.3	38	46.5	80.1	0	0	0	7	147
18	2378	28	41.1	22	38.2	71.4	0	0	1	2	67
19	4478	26	43.8	37	38.8	81.5	0	0	1	8	111
21	4501	31	42.9	36	44.9	82.3	0	0	0	14	103
22	3785	27	35.9	34	44.2	39.3	0	0	1	9	97
23	4379	21	28.5	18	43.4	81.2	0	0	0	4	49
24	18	27	33.7	27	47.1	91.3	0	1	0	6	61
25	1525	30	44.3	34	47.3	37.2	0	0	0	7	110
26	4318	21	51.3	36	33.6	51.2	0	0	0	11	98
27	1900	29	37.6	32	41.5	35.4	0	0	0	10	82
28	2038	24	42.7	32	46.9	37.5	0	0	0	5	83
29	4032	21	11.3	13	29	47.5	0	0	0	5	40
30	2927	31	45.2	37	45.3	70.3	0	0	1	10	110
32	361	32	31	28	46.4	56.6	0	0	1	6	75

Name	Type	Size	Value
data	DataFrame	(185, 411)	Column1 player nationality position squad age
XGK	DataFrame	(185, 10)	age psxg_gk games_starts passes_pct_launche
XGK_test	DataFrame	(37, 10)	age psxg_gk games_starts passes_pct_launche
XGK_train	DataFrame	(148, 10)	age psxg_gk games_starts passes_pct_launche
yGK	Series	(185,)	18 Overvalued 87 Overvalued 92 Undervalued

TESTING – MACHINE LEARNING - OVER VS UNDER VALUE

Test 17 – Output: Neural Network Predictions



	index	Over/undervalued
0	18	Overvalued
1	87	Overvalued
2	92	Undervalued
3	116	Overvalued
4	127	Undervalued
5	143	Undervalued
6	157	Undervalued
7	184	Undervalued
8	206	Overvalued
9	230	Undervalued
10	266	Overvalued
11	286	Undervalued
12	316	Undervalued
13	320	Overvalued
14	349	Undervalued
15	361	Overvalued
16	394	Undervalued
17	419	Undervalued
18	453	Undervalued
19	480	Undervalued
20	499	Undervalued
21	513	Overvalued
22	526	Overvalued
23	545	Undervalued

Name	▲ Type	Size	Value
data	DataFrame	(185, 411)	Column1 player nationality position squad age \ 18 1871.0 Fer <...> rvalued 6332 1 0 1.001003e+07 Undervalued [185 rows x 411 columns]
XGK	DataFrame	(185, 10)	age psxg_gk games_starts passes_pct_launched_gk \ 18 27.0 33.7 27.0 <...> 8.0 98.0 5553 10.0 121.0 6332 14.0 93.0 [185 rows x 10 columns]
XGK_test	DataFrame	(37, 10)	age psxg_gk games_starts passes_pct_launched_gk \ 116 33.0 32.6 20.0 <...> 11.0 107.0 4593 10.0 109.0 3499 8.0 96.0 3690 10.0 96.0
XGK_train	DataFrame	(148, 10)	age psxg_gk games_starts passes_pct_launched_gk \ 1096 29.0 34.6 29.0 <...> 5.0 55.0 920 15.0 70.0 1271 10.0 92.0 [148 rows x 10 columns]
yGK	Series	(185,)	18 Overvalued 87 Overvalued 92 Undervalued 116 Overvalued 127 Undervalu <...> valued 5553 Overvalued 6332 Undervalued Name: Over/undervalued, Length: 185, dtype: object
yGK_test	Series	(37,)	116 Overvalued 2231 Overvalued 419 Undervalued 750 Overvalued 4445 Undervalu <...> 4593 Overvalued 3499 Undervalued 3690 Overvalued Name: Over/undervalued, dtype: obje
yGK_train	Series	(148,)	1096 Overvalued 4626 Undervalued 2848 Overvalued 1620 Overvalued 3316 Overvalu <...> valued 920 Overvalued 1271 Undervalued Name: Over/undervalued, Length: 148, dtype:

TESTING – DATA ANALYTICS REPORT

4. Report – Data Analytics

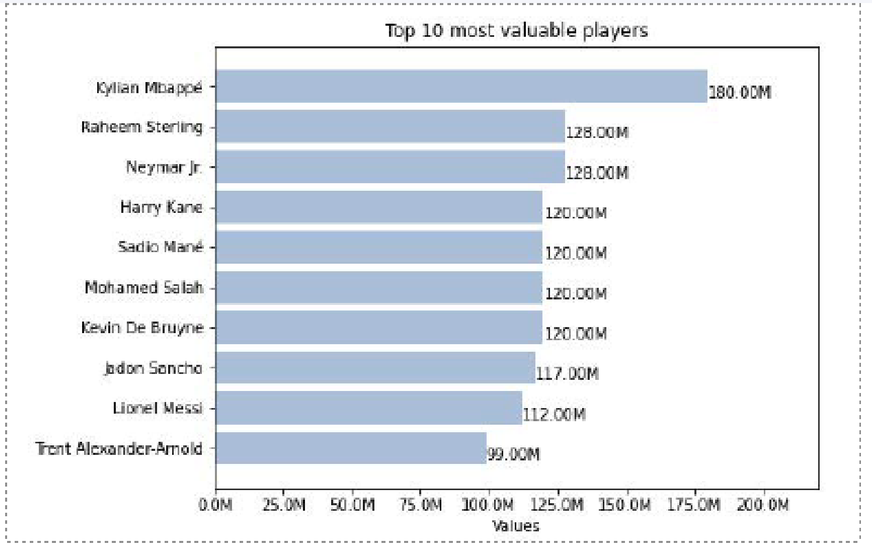
The combination of all the data scraping, analytics and machine learning is the ‘Data Analytics Report’

Test	Objective	Process	Expected Output	Actual Output	Result
18	Produce a comprehensive report for the user showing key data analysis, player comparison and the machine learning transfer market value.	Combine all the data analysis and machine learning into a report that the user can download from the website	Professional report showing the user key data analysis, player comparison and the machine learning transfer market value.	Professional report showing the user key data analysis, player comparison and the machine learning transfer market value. See following pages from the report	Pass

DATA ANALYTICS REPORT

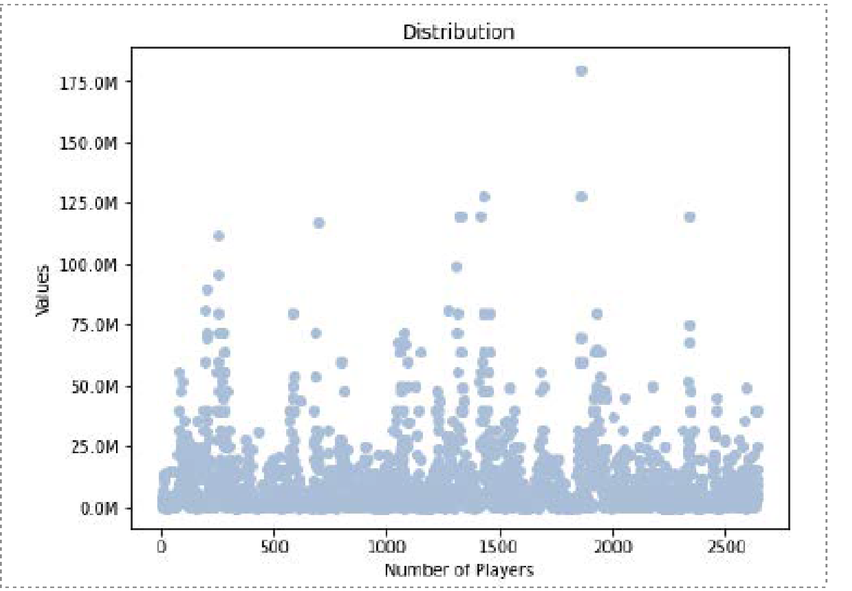
ai.football

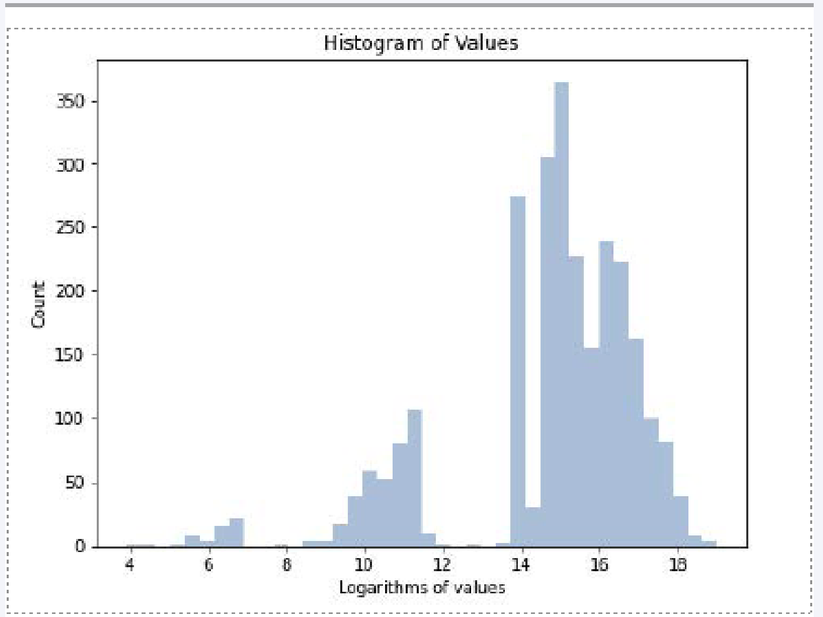
Goal! Using Machine Learning
To Build Better Teams

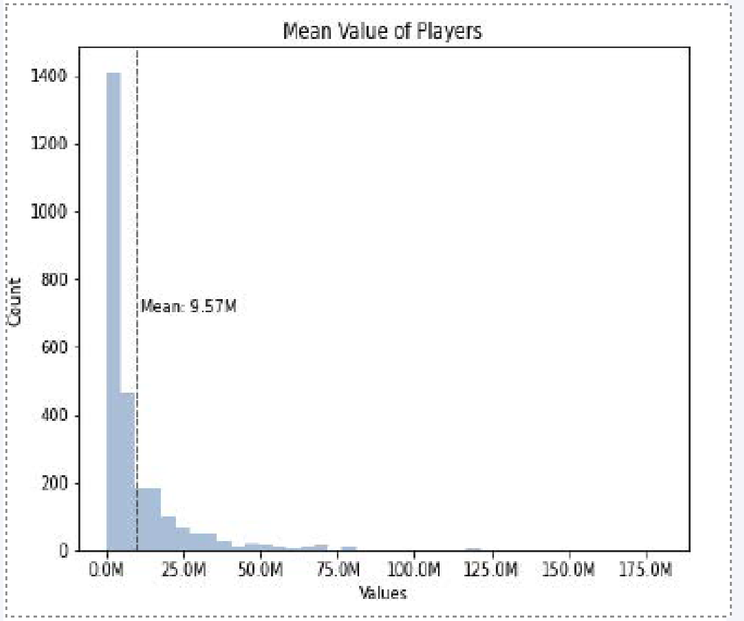


ai.football

Season 2021





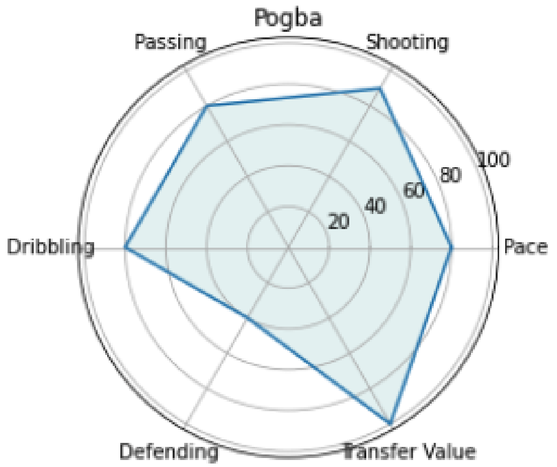


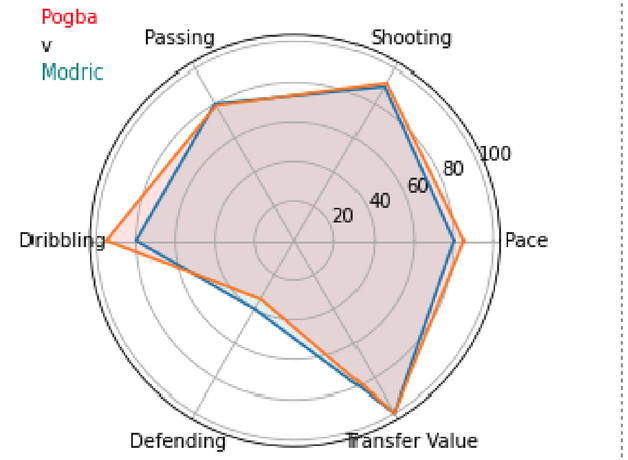
ai.football

Paul Pogba

Modric

	Pogba	Modric
Dribbling	95	80
Passing	79	80
Shooting	92	90
Pace	85	80
Defending	34	40
Transfer Value EUR	100m	94m





ai.football

A.I ANALYSIS

Name	Paul Pogba	Luka Modric
Positions	CM, DM, AM	CM, AM, DM
Age	28	36
Rating	84	82
Goals	28	22
Height	191 cm	172 cm
Preferred foot	Right	Right
Premier League	Yes	Yes
Current Transfer Value	100m	94m
A.I Value	98m	95m
Over / Under Value	Overvalue	Undervalue

Name	XG Per90	Minute	Games	Passes	Dribble
Paul Pogba ('20/21)	0.21	2045	31	428	154
Luka Modric ('20/21)	0.14	1145	25	496	102

Statisctics		
League	England. Premier League ('20/21)	Spain. Primera Division ('20/21)
Passing	79	80
Shooting	92	90
Dribbling	95	80
Pace	85	80
Defending	34	40

→ Paul Pogba transfer value is 100m but our model estimates value to be 98m

→ Luka Modric transfer value is 94m but our model estimates value to be 95m

TESTING

WEBSITE & DATABASE TESTING

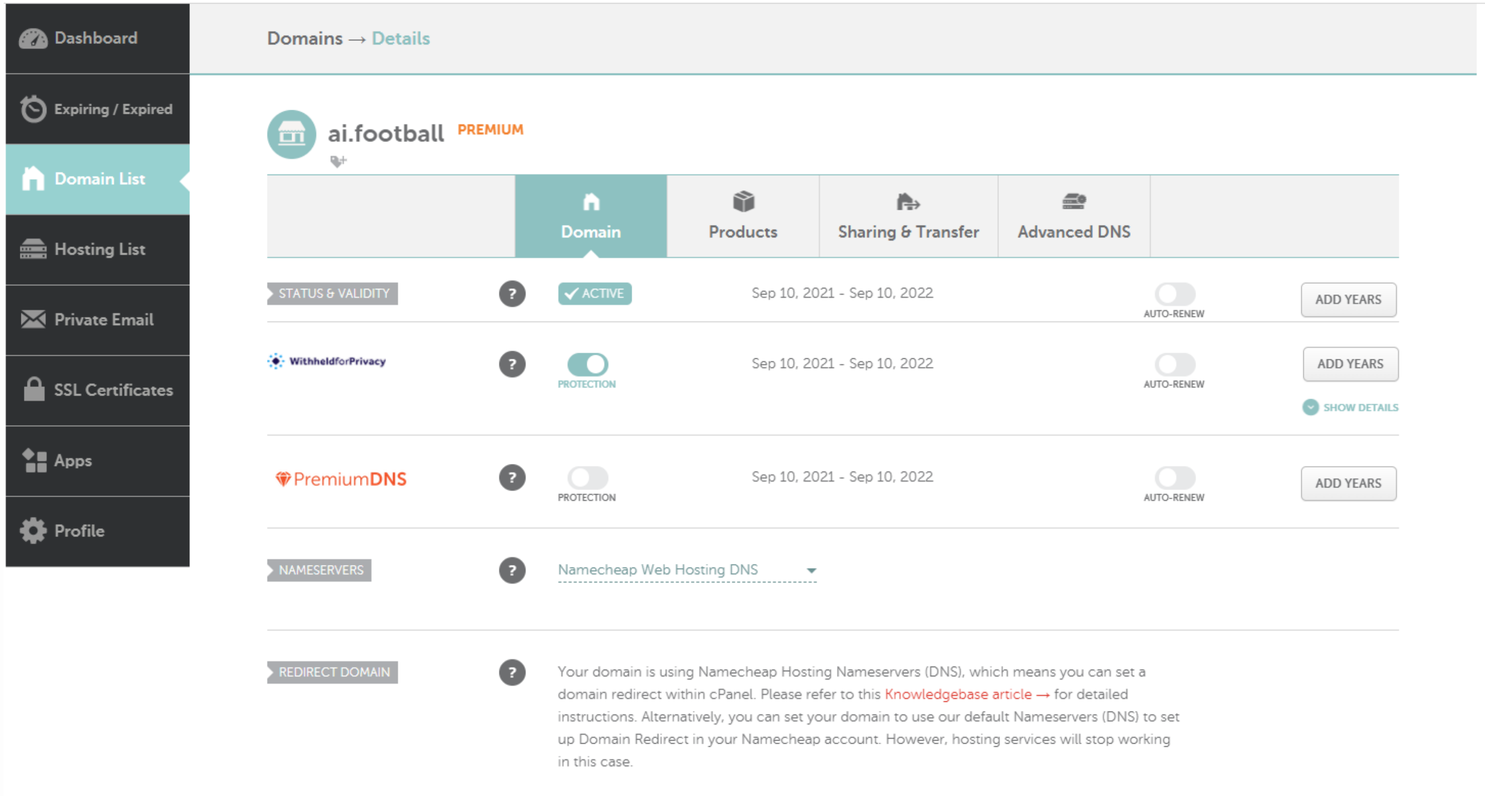
TESTING – WEBSITE DOMAIN

4. Website – Domain Registration & Hosting

The website is the user gateway to my system. The testing focuses on the creation of the website and the user functionality.

Test	Objective	Process	Expected Output	Actual Output	Result
19	Source a unique domain that would capture the target audience and show what my system is about.	Search for suitable domains	Suitable domain	Ai.football This domain says all what I want to achieve in a unique URL	Pass
20	Website hosting – now that I have purchased the domain name I need to purchase a domain hosting package with a leading domain hosting company	I read several reviews of domain hosting companies and selected one that had a 99.99% up time.	Leading domain hosting company	I purchased the hosting package to host my domain	Pass

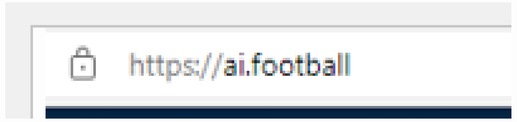
Test 19 & 20 – Output: Domain registration and hosting

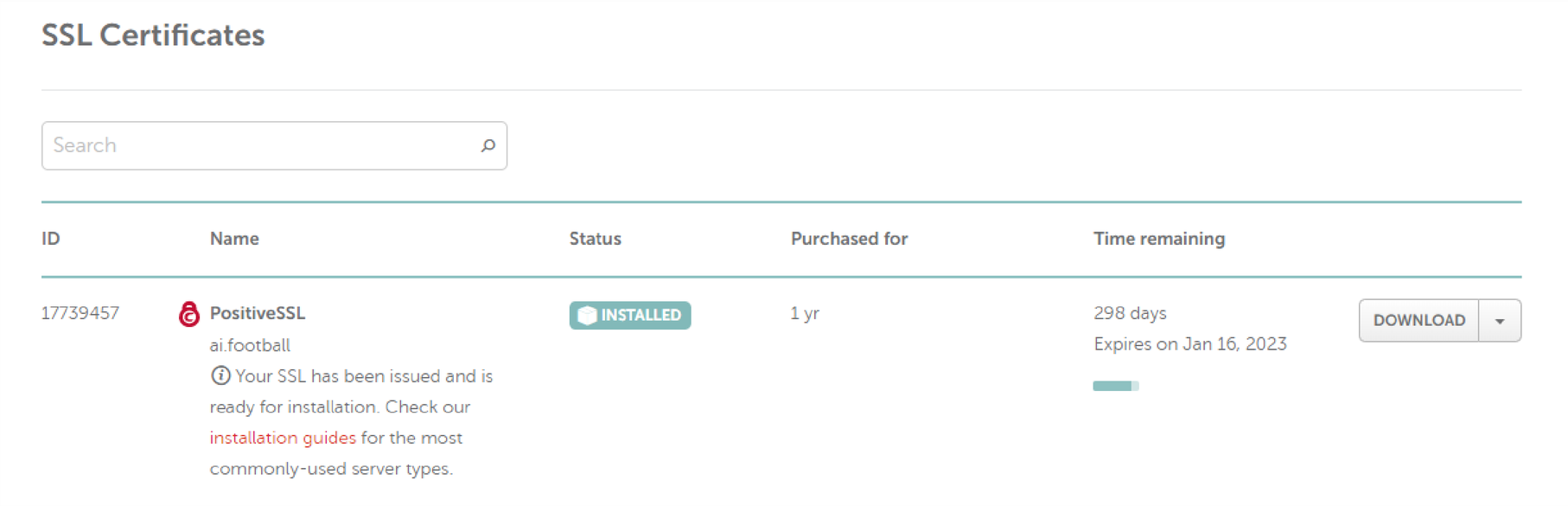


TESTING – WEBSITE SECURITY

Test	Objective	Process	Expected Output	Actual Output	Result
21	Setup a SSL Certificate to make sure that the domain has a https:// URL	Setup the domain security, configure the SSL and install in domain	That the domain loads with https://	https://ai.football loads.	Pass

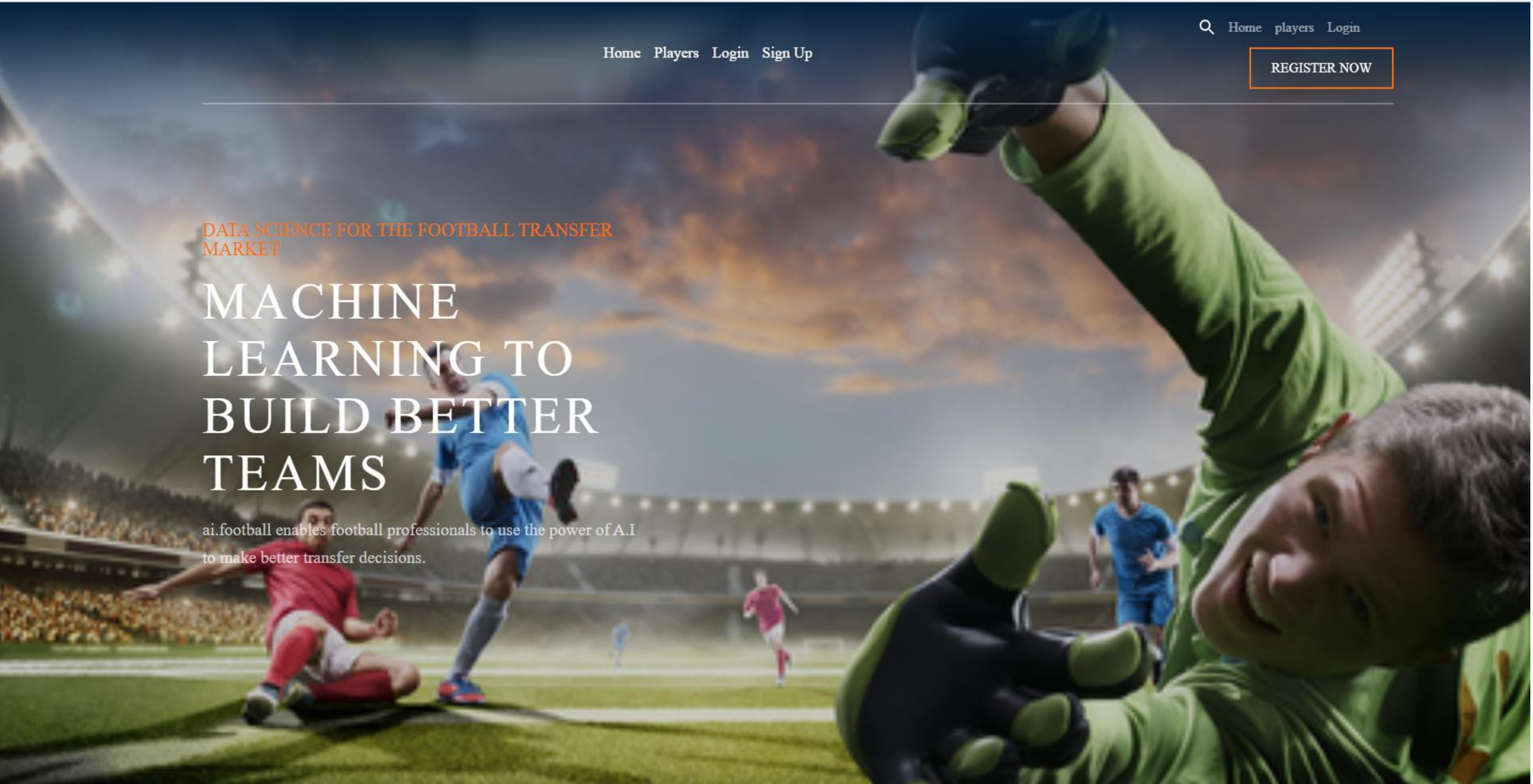
Test 21 – Output: Domain SSL & https://





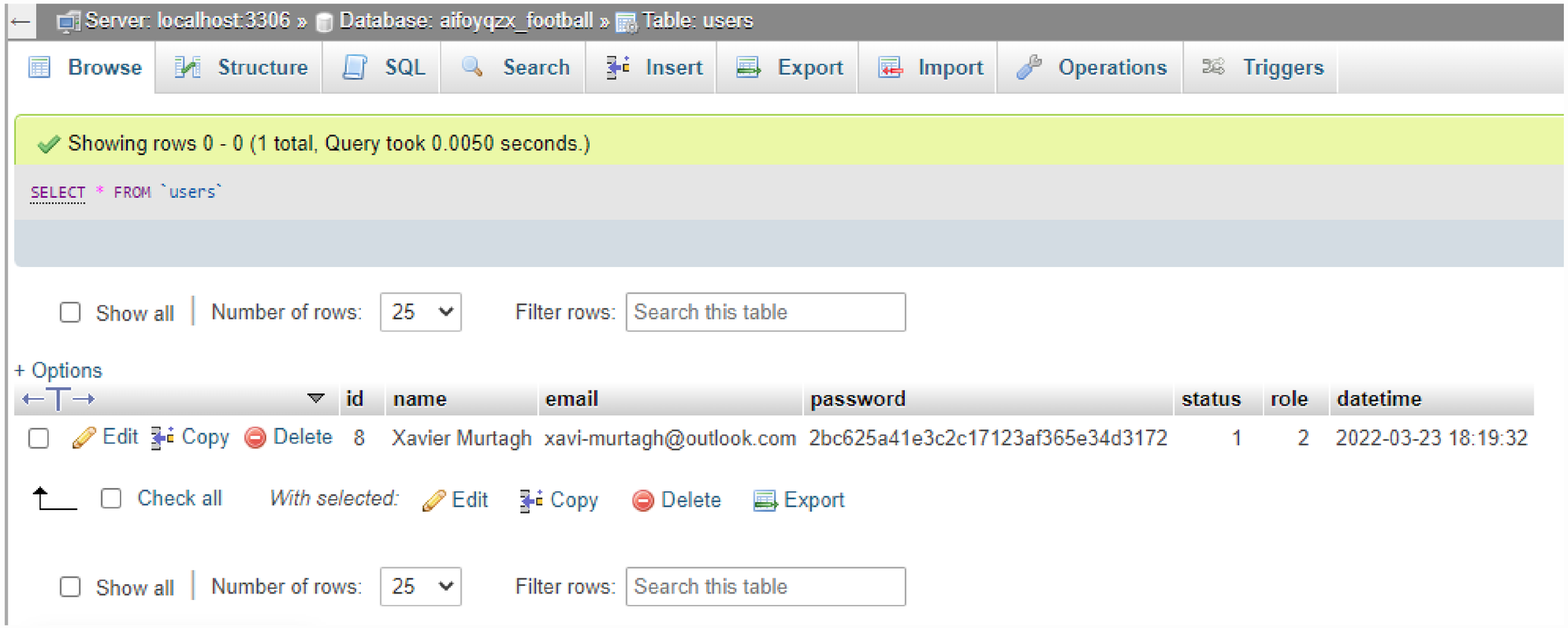
TESTING – WEBSITE FRONT PAGE

Test	Objective	Process	Expected Output	Actual Output	Result
22	User goes to ai.football and opens the front page	Index.php loads with the front page images and menu options	Frontpage opens when user types ai.football URL	https://ai.football See test output	Pass
23	Registration - user wants to register to access the system	User clicks the on the Register menu option	The registration page opens. The user then enters their details: 1. Name 2. Email 3. Password 4. Confirm password	The registration page opens. The details of the user are stored on MySQL: 1. Unique user ID 2. Name 3. Email 4. Password encrypted	Pass


ai.football
https://ai.football
Home players Login
Home Players Login Sign Up
REGISTER NOW
DATA SCIENCE FOR THE FOOTBALL TRANSFER MARKET
MACHINE LEARNING TO BUILD BETTER TEAMS
ai.football enables football professionals to use the power of A.I to make better transfer decisions.

TESTING – WEBSITE USER REGISTRATION

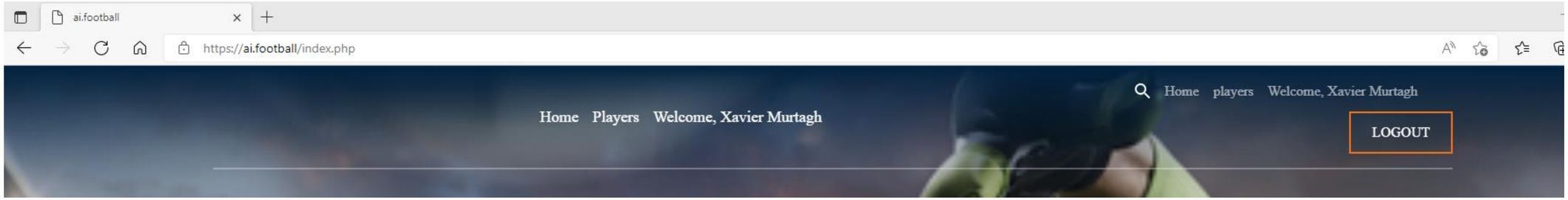
Test 23 – Output: User registration

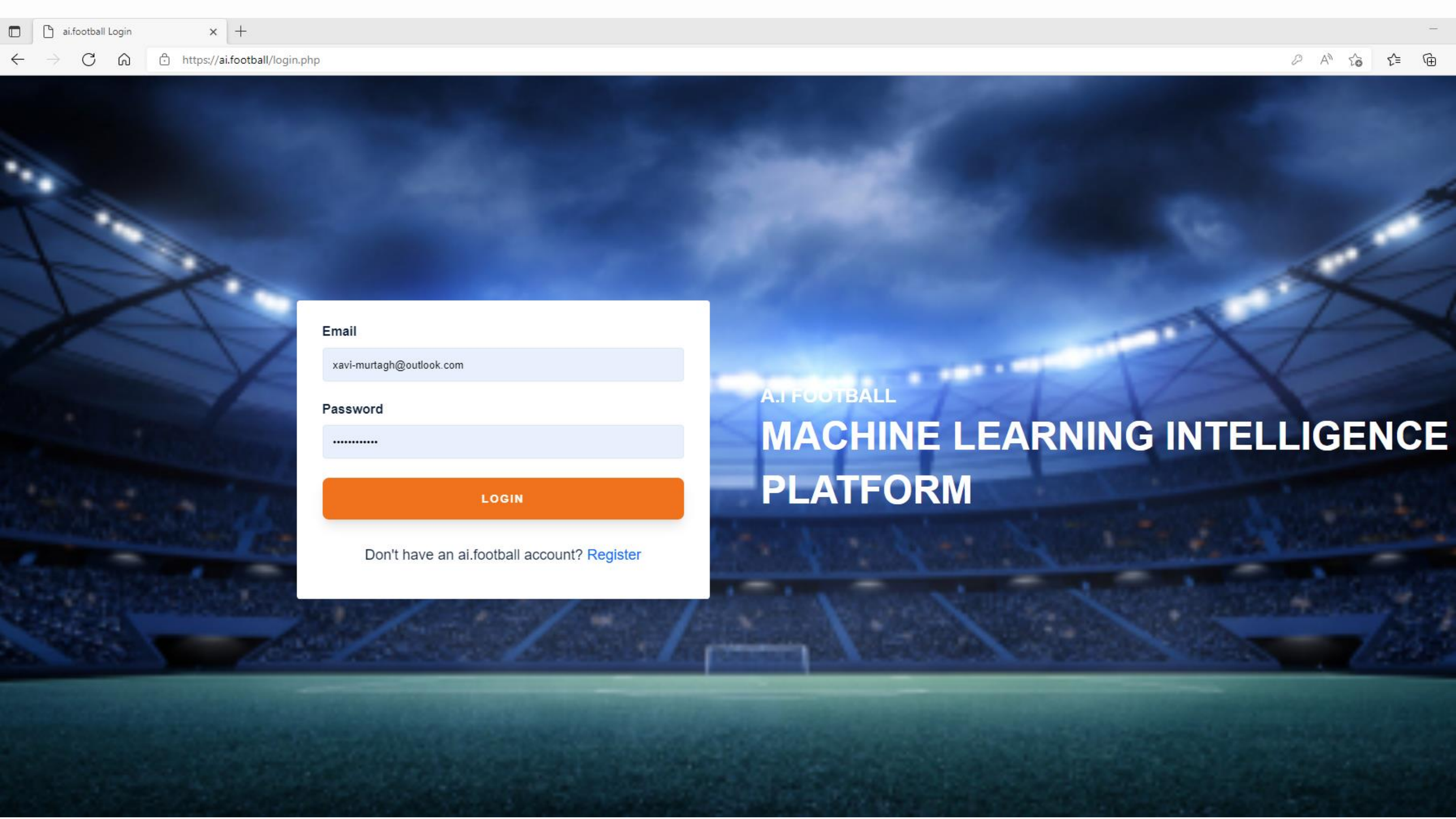


TESTING – WEBSITE LOGIN

Test	Objective	Process	Expected Output	Actual Output	Result
24	Test login – this is to test whether a user who previously registered is able to login again. When the user does login their name should be shown on the menu.	Login details are checked against those stored in the user registration. If match the page: http://ai.football/player.php is accessible	Welcome {user_ID} http://ai.football/player.php is accessible	“Welcome Xavier Murtagh” http://ai.football/player.php is accessible	Pass

Test 24 – Output: User login

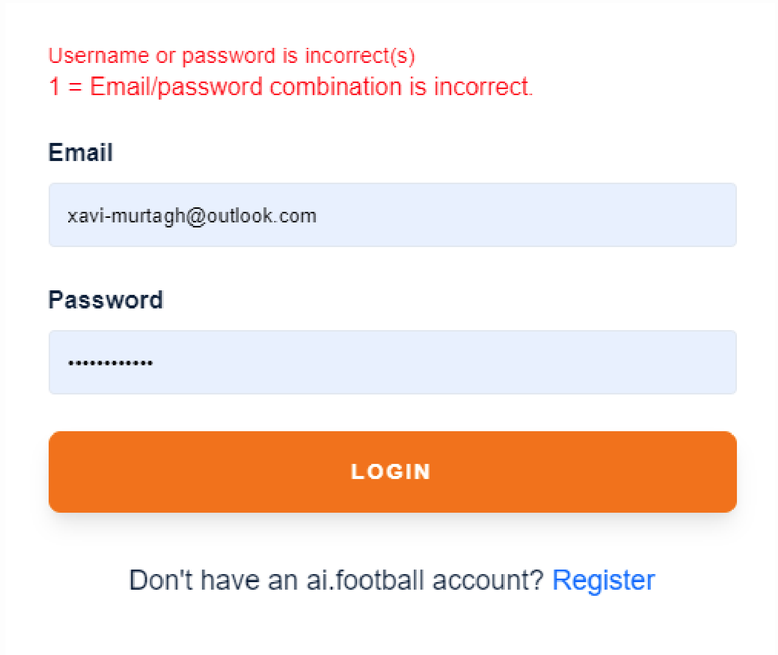



ai.football Login
https://ai.football/login.php
Email
xavi-murtagh@outlook.com
Password
LOGIN
Don't have an ai.football account? Register
A.I FOOTBALL
MACHINE LEARNING INTELLIGENCE
PLATFORM

TESTING – WEBSITE LOGIN

Test	Objective	Process	Expected Output	Actual Output	Result
25	Test incorrect login. This test will show what the output is when the user tries to enter an incorrect password	Username and password are checked and if they do not match an error is shown asking the user to recheck their login details	Error message stating that the password does not match or the email ID is already registered	Error message stating that the password does not match or the email ID is already registered	Pass

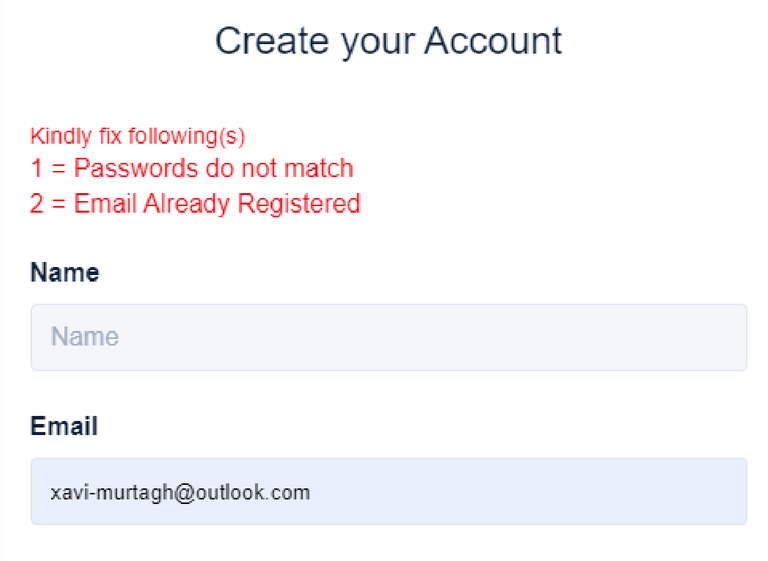
Test 25 – Output: Incorrect User login



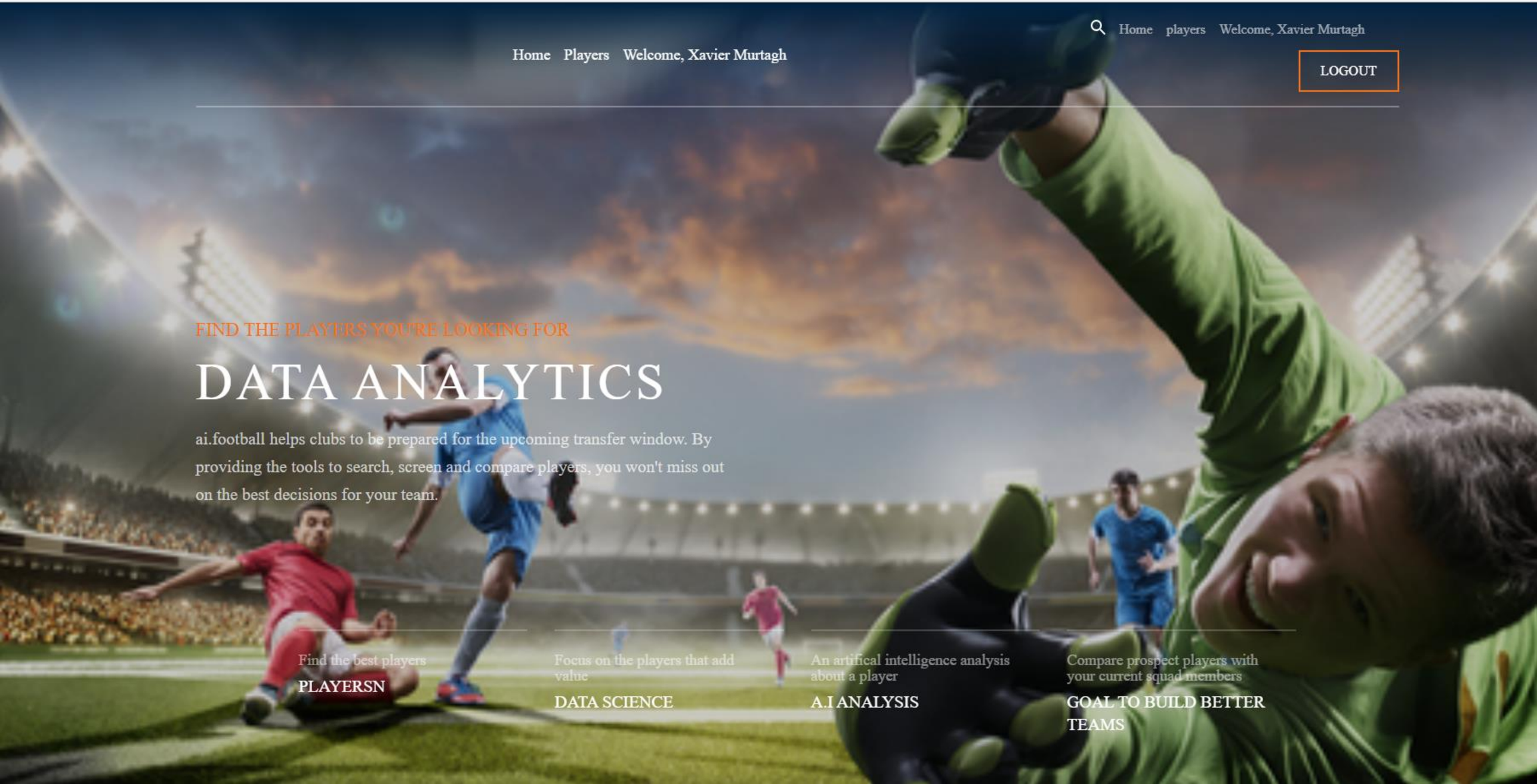
TESTING – WEBSITE ACCOUNT

Test	Objective	Process	Expected Output	Actual Output	Result
26	Test incorrect account. This test will show what happens when the user creating an account tries to enter unmatching passwords and uses an email that's already in use	Username and password are checked and if they do not match an error is shown asking the user to recheck their login details	Error message stating that the password does not match or the email ID is already registered	Error message stating that the password does not match or the email ID is already registered	Pass

Test 26 – Output: Incorrect Account



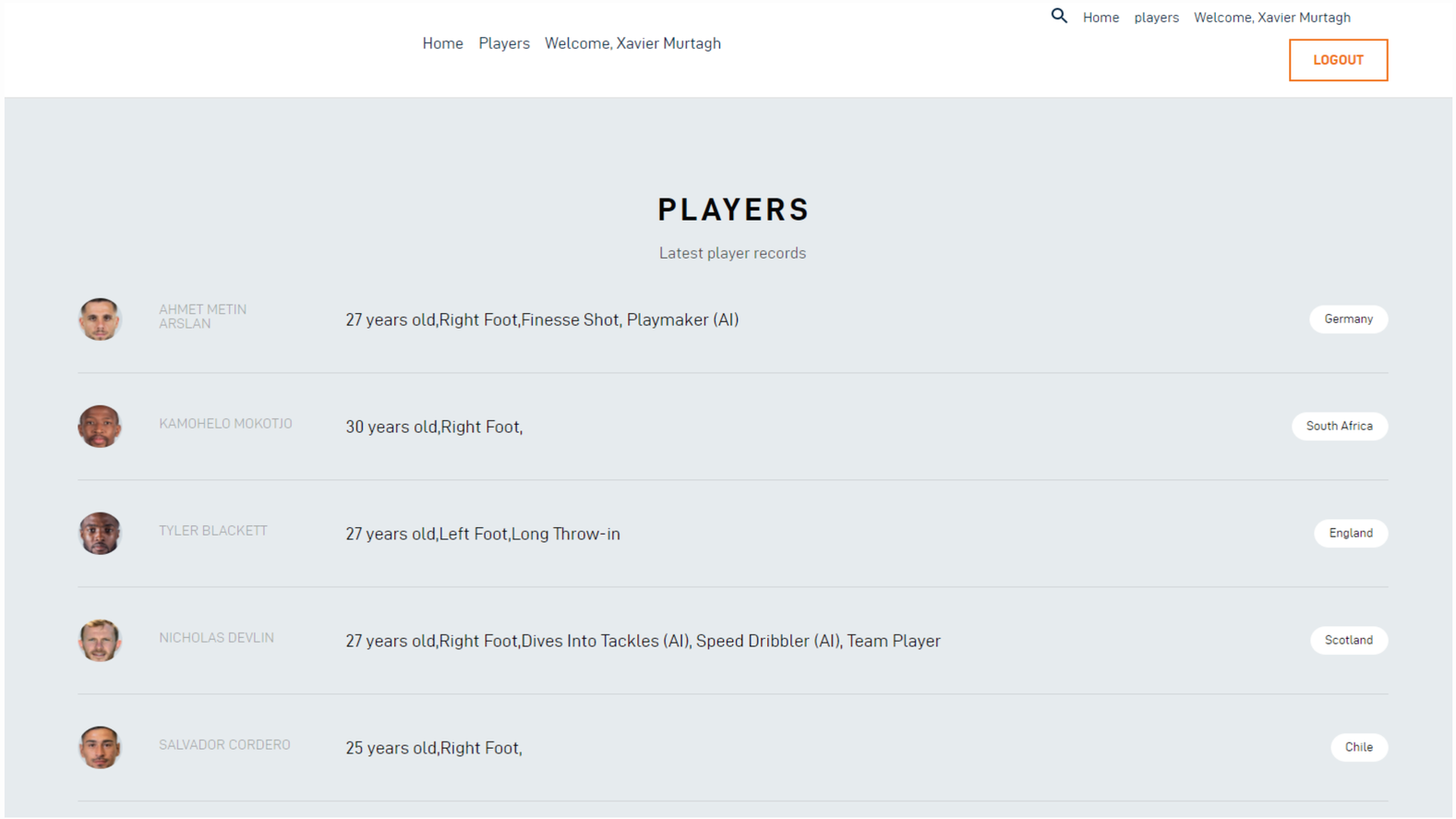
Test	Objective	Process	Expected Output	Actual Output	Result
27	Data Analytics page	Once user is logged in they can access the data analytics page	The data analytics page loads	The data analytics page opens See following page.	Pass


FootBall
https://ai.football/players.php
Home Players Welcome, Xavier Murtagh
Home players Welcome, Xavier Murtagh
LOGOUT
FIND THE PLAYERS YOU'RE LOOKING FOR
DATA ANALYTICS
ai.football helps clubs to be prepared for the upcoming transfer window. By providing the tools to search, screen and compare players, you won't miss out on the best decisions for your team.
Find the best players
PLAYERSN
Focus on the players that add value
DATA SCIENCE
An artifical intelligence analysis about a player
A.I ANALYSIS
Compare prospect players with your current squad members
GOAL TO BUILD BETTER TEAMS

TESTING – WEBSITE PLAYERS LIST

Test	Objective	Process	Expected Output	Actual Output	Result
28	Player page – registered user is able to access the list of players. The user can see the key abilities of the players and can click on a player for more information.	Only after a successful login is the user able to access the players page. It has a list of players that the user can scroll through	Load players.php page and display the player details	See below – the page successfully loads	Pass

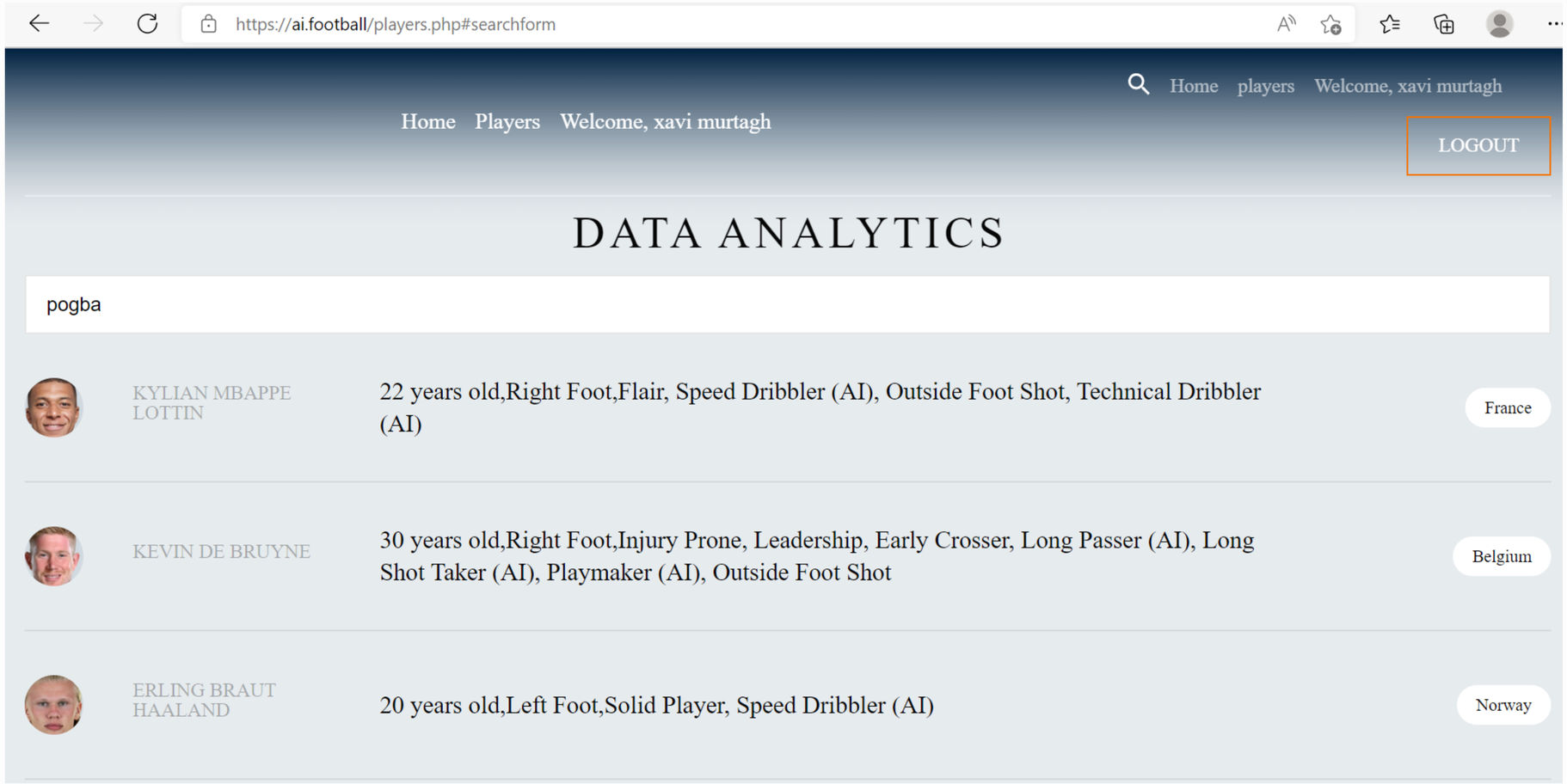
Test 28 – Output: Players Page



TESTING – WEBSITE PLAYER SEARCH

Test	Objective	Process	Expected Output	Actual Output	Result
29	Registered user wants to search for a particular player	Search for the player name	The player's page loads	The page for the player loads	Pass

Test 29 – Output: Player search



https://ai.football/player-details.php?id=195864

Home Players Welcome, Xavier Murtagh

Home players Welcome, Xavier Murtagh

LOGOUT

PAUL POGBA

PLAYER ANALYSIS

Here are the statistics of Paul Pogba. ai.football uses the power of artificial intelligence to evaluate a player's transfer value to determine if they are over or undervalued.

PAUL POGBA

TESTING – WEBSITE A.I VALUATION

Test	Objective	Process	Expected Output	Actual Output	Result
30	The final and the most important test is to display a player valuation and compare this to the 'predsOLS' which is the predicted value from neural network algorithm.	This test takes the player ID from the user search and loads the details of the player. It takes the 'predsOLS' field and populates the 'A.I Value' field The field 'Potential Savings' is the difference between the current transfer value and 'A.I Value'.	1. Player page loads 2. Player metrics are loaded from the database 3. A.I Value field is populated from the database 4. The transfer value is loaded 5. The A.I Value is 'predsOLS' field 6. The Potential Savings field is the difference between the transfer value as and the machine learning derived valuation	1. Player page loads - Yes 2. Player metrics are loaded from the database - Yes 3. A.I Value field is populated from the database - Yes 4. The transfer value is loaded - Yes 5. The A.I Value is 'predsOLS' field - Yes 6. The Potential Savings field is the difference between the transfer value as and the machine learning derived valuation - Yes	Pass

Test 30 – Output: A.I Valuation

PAUL POGBA

ai.football uses the power of artifical intelligence to understand what drives player valuation and how clubs can deploy this to build better teams.

- Date of birth: 15/03/1993
- Age: 28
- Rating: 84
- Height: 191
- Preffered Foot: Right
- Skills: Injury Prone, Flair, Long Passer (AI), Long Shot Taker (AI), Playmaker (AI), Outside Foot Shot, Technical Dribbler (AI)

Goal! Using A.I to Build Better Teams

Overall A.I RATING	Player TRANSFER VALUE	Player A.I VALUE	Player POTENTIAL SAVINGS
87	10000000	98129368	1870632
ai.football	ai.football	ai.football	ai.football

TESTING – YOUTUBE TUTORIAL & TESTING VIDEO

Objective – Youtube Tutorial & Testing

www.ai.football is the combination of a lot research and development.

I wanted to create a tutorial for the users to be able to navigate the website and all the functions. The video also confirms all the output of the 30 tests that I have listed so confirm that the system does what I set out to do.

The url of the youtube video is: https://youtu.be/ZJkUGecdgUQ

EVALUATION

EVALUATION OF OBJECTIVES

EVALUATION

Introduction

Whilst researching this project, I read the following article by Arsene Wenger. The former Arsenal manager effectively admitted that football is unpredictable https://www.irishexaminer.com/sport/soccer/arid-20321805.html.

Football's beauty is its unpredictability, says Arsene Wenger

Arsene Wenger says his undiminished, ever-lasting love of football is based on the game's glorious unpredictability.

THU, 02 APR, 2015 - 01:00

DARREN NORRIS

The veteran Arsenal boss admits that uncertainty gives him pre-match jitters but claims the game wouldn't be so universally popular if it was logical.

"Of course I get nervous before the games because football is not mathematics," Wenger said. "In maths every day you know that one plus one is two. In football, one player plus one player doesn't always add up to two players.

There are no certainties in anything and certainly not in sport where there are so many variables. However, as I identified in the Analysis section, my system is designed to provide a way to help football clubs make better decisions. In the following evaluation section, I wanted to highlight the limitations of the system and things I want to improve in future updates. I also want to see whether the objectives I set out in the Analysis section my project were met.

EVALUATION – DATA LIMITATIONS & FUTURE VERSIONS

Data Outliers

In my analysis of the project, I highlighted one factor that could influence the results. The data used in the 2020 football season was influenced by the COVID-19 pandemic. Therefore, it is likely that this will have resulted in some misleading data and statistics. However, this also identifies one important point about machine learning. There will always be outliers and data points that do not fit into the perfect model. Therefore, it is important to point out that even though the 2020 season is a data set where the transfer value of the players was affected, it is still an important data set.

Choice of Leagues

It is difficult to compare the leagues from different countries. I chose the top leagues in Europe, but there are many other leagues where the data could determine the value of a player.

Female Players

As I used data from the top European male leagues, I have excluded female football players from my analysis. This wasn't through choice, but because the data set for female players is currently smaller and harder to obtain. When I develop this system further, I will try and source data from the female leagues throughout Europe.

Injuries

One of the problems of my model is that it doesn't look at injuries. The dataset does not have an injury column. Some injuries are temporary, but some are more permanent. An injury will have an impact on a player's value and the type of injury will have an even greater impact.

Factors Influencing Player Values

In my system I assume that the value of a player is determined by a particular set of statistics. I have shown that these metrics do have an influence on player values and can be predictive. However, there are many other factors that will influence the value of a player. A future version would expand the number of player attributes that would be used in the regression model.

EVALUATION – DATA LIMITATIONS & FUTURE VERSIONS

Actual Transfer Values

The market value that I have used to compare the expected market value and the market value is from the website Transfermarkt. In a future version I will assess alternative sources for this data to ensure that there is a benchmark. For example, it would be more accurate to have had the actual transaction value the club paid as we do not know if a club bought a player for more or less than what was reported. For example, in my research I found that the player Eduardo Camavinga was sold for £28 million despite having a market value of £50 million.

Playstyle

All teams have different style of play, and this means that not everyone will work well in a team that need a different skillset to the one the player has. Let's use Burnley as an example: Burnley like to hit long balls to their tall attackers where they win the ball and score and because of this their new striker Wout Weghorst has scored goals and increased his transfer value. However, if Lionel Messi were to move to Burnley, it is unlikely that he would have much effect on matches due to his small frame so while he may be better value for money than Wout Weghorst, Wout Weghorst would be a much better player for Burnley for the price. This is something I could develop in future versions if I found statistics about teams.

Medical history

Medical history of a player is scrutinized by football teams, and we see this now with the medical tests that must be done just before completing a transfer. More teams are wary of players injury history before signing them. Injuries can be extremely harmful for football clubs and can change a team's season due to the extended period on the sidelines. Liverpool for example last year lost many of their top players and had to rely on youth players to play which meant that they dropped from the title challenge. Players who have been plagued by injury are also seeing their market value go down due to this and it is something that I would want to add in a future development of ai.football as it has led to many promising players to having disappointing football careers.

EVALUATION – TECHNICAL LIMITATIONS & FUTURE VERSIONS

Data

The dataset used is a ‘snapshot’. It is not real time. A future version of the system would not be limited by this as it would have real time data. The user would then be able to analyze the system using current transfer market values.

Type of Machine Learning Algorithms

There are a wide variety of machine learning algorithms. The SKLearn is a very well tested neural network that has been used to solve many machine learning problems. In my evaluation I want to highlight that there are many different types of neural networks that I could have used. Each algorithm would have produced slightly different results. Each of these would have been valid but they would have produced different results.

Training Algorithm & Neural Network

In the design section I identified weights that I would apply to the neural networks for each of the three football seasons. My approach to the weights that I applied was that I chose a higher weight for the current season and the prior seasons had a lesser weight. However, the size of the weight is a factor that can be adjusted, and this will lead to different results.

Hosting

The website has been tested in the testing section and it passed all the key tests. I have also tested with several users, however there is a limitation that with all websites that if there were a lot of users at the same time there many be potentially be system issues. This is something I would try to fix in future versions.

Application to other Sports

In the Analysis section of this project, I identified that other sports could use my machine learning approach to identify those players that were over or under valued relative to their transfer market value. The original statistical analysis was in baseball where Billy Beane used statistics to identify players that had fallen out of favour among the bigger baseball clubs, but who could still deliver. My system is built for football players, but it could also be used in other sports such as cricket, rugby etc. if statistics for these sports were readily available.

EVALUATION – MEET OBJECTIVES

Meet Objectives

In the Analysis section I set out several key objectives that I wanted to achieve in this system. I therefore want to see if each of them has been achieved.

Objective - research & develop an idea that would have a real word application

The research section of the project has numerous stories about the football transfer market and the vast amount of money paid for players. The headlines clearly show that this is not always well spent. Many players fail to live up to the hype. It was clear to me that some clubs bought players because they were in good form. They may have scored some goals recently, but there wasn't a longer assessment to see how the player has played over multiple seasons. If there had been a more data driven evaluation of the player's performance, then the club could have saved money.

Objective – develop a website that was modern and professional

In the research stage I also found other websites with details on football players, but these were limited. There isn't a consolidated a data set that I could use to build a comprehensive set of data analytics. As shown in the testing and the system I was able to build this using data scraping from multiple websites.

Objective – create a data set across multiple leagues and seasons

In the research stage I also found other websites with details on football players, but these were limited. There isn't a consolidated a data set that I could use to build a comprehensive set of data analytics. As shown in the testing and the system I was able to build this using data scraping from multiple websites.

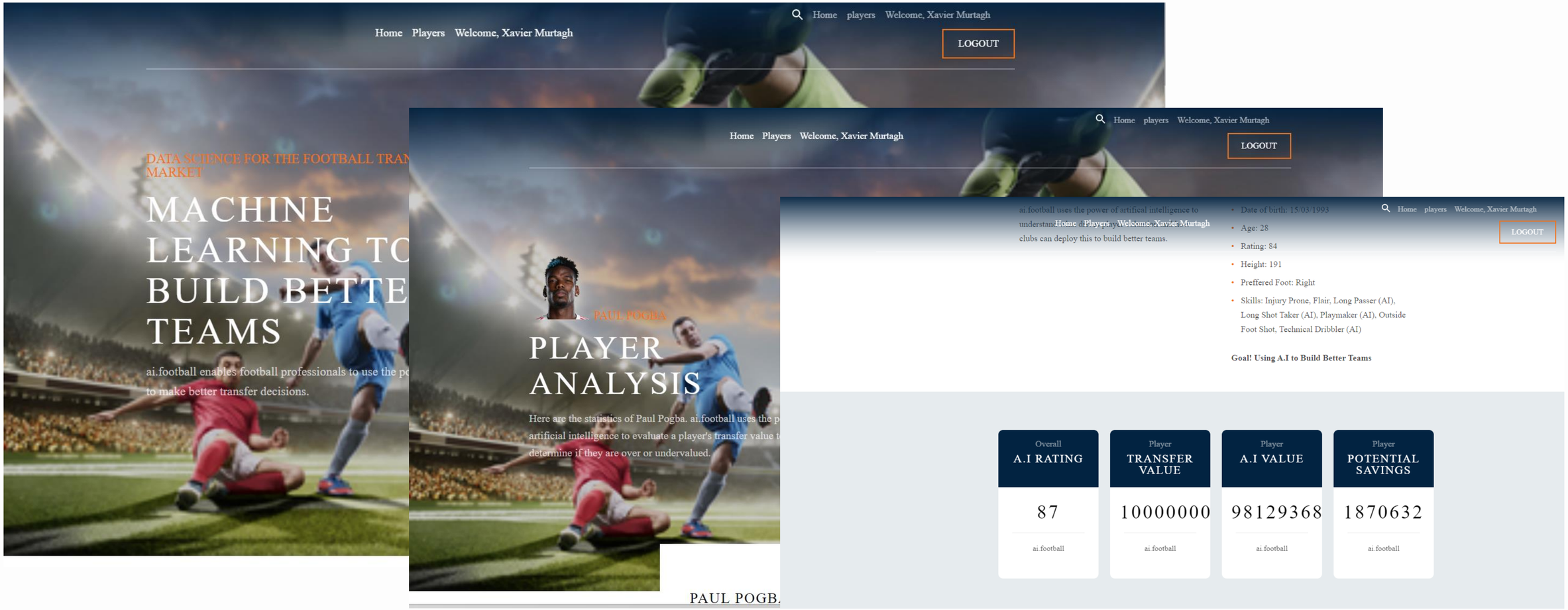
Objective – develop professional data analytics report

The output of the initial coding stage was to develop a report that the user can use to assess different player attributes. It has several graphics that the user can use to quickly identify different attributes. The report shows relevant data on the player transfer market. It then expands into a detailed player analysis where the user can compare the attributes of one player vs another. This will enable the football club make better decisions.

EVALUATION – MEET OBJECTIVES

Objective – create a professional website for the system learning algorithm to determine if a player is over or under valued

www.ai.football is the combination of a lot of elements, the data scraping, the data output, the analytics and the machine learning algorithms. The detailed testing has shown the output of all the development with result being a professional, modern website (see the following) with a unique A.I algorithm to help the club ascertain if a player is over or under valued. I have recorded a Youtube video to show the system working how I intended



EVALUATION – MEET OBJECTIVES – POST DEVELOPMENT INTERVIEW

Follow Up User Interview – Post Development

One of the most important assessments to see if I met the objectives of the project is the user assessment of the system. I therefore contacted the manager I spoke to during my research and asked him to give me feedback on the system. As the system is available on the internet, I sent him the ai.football URL and asked him to spend some time evaluating the system and website. I then followed up with the following interview.

Club Manger Evaluation of the System Interview

Q. Thank you for taking time to speak to me again and to evaluate the system. I would like to get your initial thoughts on it

“It is very good. I like what you have done. When we spoke before I could understand what you wanted to build and the problem you were trying to solve, but I had no idea how you would make it into a website.

Q. Can you give me your assessment of the website

“It was easy to navigate and to use. The front page was simple, but it clearly explains in a few words what the site is about. I was able to create a login from my email address. From there I could access the page about the different players. I was able to scroll through the page of players. There is a a lot of comprehensive set of players that I could search for”

Q. Did you find the player analytics useful?

“Yes, it was. To have all the information of so many players is very useful. I could see myself making use of this when we are reviewing players that we would be interested in buying for our club”.

EVALUATION – MEET OBJECTIVES – POST DEVELOPMENT INTERVIEW

Q. Did you download the player report

“I did. It was very good. It had lots of information on player values and the graphs that were very helpful in putting everything into context. There are so many statistics in football to simplify. As they say, a picture paints a thousand words”

Q. In the next part of the interview I want to ask your opinion on the artificial intelligence predicted value analysis?

“I have no idea how A.I works, but I would like to learn more about it! To see it applied to a player’s transfer value was very interesting. I did a search for a number of players and looked at what the A.I Value was compared to the transfer value. The results were fascinating. There were some players that the website stated were over valued, but what particularly interested me was those that it said were undervalued. As a smaller club this is what we are always trying to do. We need to buy players that give us the best value relative to their performance. The big clubs can afford to spend what they like, but can’t afford that luxury. We’d like to be the Moneyball of football and look at Billy Beane who revolutionized the baseball transfer market”

Q. Do you have any feedback about the system and what features would be useful in future versions?

“I know there is a lot of data in your system, but it would be very useful to have more players from the lower leagues. It would also be helpful to have the player details for the current season. The transfer window is only open twice a year but we spend a lot of time throughout the year looking at players. Our scouting team would find this web site really useful if it had lots of current info”.

Q. Any final comments?

“This a great site and I can see this developing into something that all clubs could use. Data science is growing exponentially in all sports, but particularly football. Clubs are now businesses and are increasingly ran as such. The amount of money in player transfers is staggering. Your idea of using artificial intelligence to help clubs like ours make better transfer decisions could be a real game changer. I also wanted to say that the name of your website, “ai.football” is a great name!

APPENDIX

APPENDIX & BIBLOGRAPHY

APPENDIX & BIBLIOGRAPHY & DATA SOURCES – MACHINE LEARNING

1. J. VanderPlas. (2016), “Python Data Science Handbook”
2. I. Hendriks (2017), “Modelling the transfer prices of football players”
3. P.W. Holland, R.E. Welsch (1977), “Robustness regression using iteratively reweighted least-squares”
4. https://res.mdpi.com/d_attachment/entropy/entropy-23-00090/article_deploy/entropy-23-00090-v3.pdf
5. https://thinkml.ai/artificial-intelligence-ai-in-football-soccer/
6. www.statista.com/statistics/261223/european-soccer-market-total-revenue/
7. https://www.linkedin.com/pulse/moneyball-why-book-which-now-17-years-old-more-than-palser-thorne
8. https://atrium.ai/resources/moneyball-meets-ai-transforming-business-with-data-driven-insights/
9. https://medium.com/swlh/the-holy-grail-of-sports-tech-machine-learning-moneyball-and-inspiration-1e4f472c67c9
10. https://www.forbes.com/sites/barrylibert/2018/08/31/machine-learning-is-a-moneyball-moment-for-companies/
11. https://www.skyfilabs.com/project-ideas/moneyball-sports-analyzer-using-machine-learning
12. https://thinkml.ai/artificial-intelligence-ai-in-football-soccer/
13. https://deepmind.com/blog/article/advancing-sports-analytics-through-ai
14. https://medium.com/swlh/the-holy-grail-of-sports-tech-machine-learning-moneyball-and-inspiration-1e4f472c67c9
15. https://www.forbes.com/sites/barrylibert/2018/08/31/machine-learning-is-a-moneyball-moment-for-companies/
16. https://mljar.com/machine-learning/compare-ml-algorithms/
17. https://deepmind.com/blog/article/advancing-sports-analytics-through-ai
18. https://thinkml.ai/artificial-intelligence-ai-in-football-soccer/
19. https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/
20. https://www.transfermarkt.co.uk/
21. https://www.fbref.com
22. https://www.sofifa.com
23. https://www.scisports.com/
24. https://mljar.com/machine-learning/compare-ml-algorithms/
25. https://jtablesaw.github.io/tablesaw/userguide/ml/Moneyball%20Linear%20regression.html
26. https://realpython.com/python-ai-neural-network/

APPENDIX & BIBLIOGRAPHY – HTML AND PHP

1. HTML tutorial for Beginners, https://www.youtube.com/watch?v=qz0aGYrrlhU
2. https://www.freecodecamp.org/news/html-basics-for-beginners/
3. https://developer.mozilla.org/en-US/docs/Learn/Getting_started_with_the_web/HTML_basics
4. https://www.codecademy.com/learn/learn-html
5. https://www.codecademy.com/learn/learn-php
6. https://www.codecademy.com/learn/paths/learn-how-to-build-websites
7. https://www.learn-php.org/
8. https://www.w3schools.com/html/
9. https://www.w3schools.com/php/
10. https://www.w3schools.com/howto/howto_website.asp
11. https://www.wpbeginner.com/guides/
12. Build a website Tutorial HTML Full course, https://www.youtube.com/watch?v=pQN-pnXPaVg
13. HTML for beginners freeCodeCamp, https://www.freecodecamp.org/news/html-crash-course/
14. How to build a website from scratch, https://www.youtube.com/watch?v=9vGdmI8zz0s
15. Web Development Using PHP, https://www.youtube.com/watch?v=PGvrnas2_Pk
16. PHP Programming Language Tutorial, https://www.youtube.com/watch?v=OK_JCtrrv-c&vl=en
17. https://blog.hubspot.com/website/html
18. https://websitesetup.org/