**Predicting and Analyzing the 2022 World Cup**

Michael Mallon

**GitHub URL:**

<https://github.com/michaelmallon/UCDPA_michaelmallon>

**Abstract**

Football is a multi-billion-dollar global industry and the most followed sport around with world. With over 3 billion fans universally, the sport’s speed and unpredictability allows soccer to thrive as one of the world’s most popular sports.

The financial success of the sport is driven by a range of factors, including the sale of stadium rights, player transfers, ticket sales, and sponsorships. These sources of revenue can have a significant impact on a team's financial performance. A team's ability to make smart investments in players can be a key factor in its success.

The purpose of this project was to predict and analyze the 2022 World Cup using Python programming and real-world data. The project utilizes FIFA International Team World Rankings and EA Sports FIFA's database of players, creates a new financial driven metric for player ratings to recreate and simulate the tournament. The results of the simulation are then used to gain insights and perform analysis on the data.

This project offers a unique opportunity to explore the potential outcomes of the World Cup and gain a deeper understanding of the participating teams and players.

# **Introduction:**

Data Analytics is an important tool in the world of football, making data-driven decisions that can improve performance and drive financial success. By analyzing data on player performance, team tactics and market trends, clubs can identify opportunities to improve their operations and invest in players that are likely to generate a positive return on investment.

The role of performance analysis within football is more important than ever. Whether it’s the opposition, potential transfer targets or last weekend’s fixture, analyzing performances and data can be the difference between success and failure.1

With the ongoing 2022 World Cup in Qatar also known as ‘The Most Expensive World Cup in History’ grossing over an estimated $200 billion spend this project seems a fitting and unique way to explore the capabilities of data mining and the analytical methods Python holds.

This project utilizes the EA Sports FIFA 23 and FIFA World Team datasets, providing a wealth of information with 80+ datapoints on players and their international teams. Using this data this project simulates the 2022 World Cup using a custom built python function. This function plays each team’s corresponding attack, midfield and defense against each other to generate the outcome of a match.

Teams ratings are data mined using off a normalized weighted value using five key values in a player and a team:

* EA Sports FIFA Player:
  + Overall Rating
  + Potential Rating
  + Value in Euro
  + Wages in Euro
* Official FIFA International Team World rankings

Team form is then created by increasing and decreasing the normalized team rating dependent on the outcome of the match.

In-depth analysis is then performed on the resulting dataset of players and their international teams.

# **Datasets:**

## [EA Sports FIFA 23 Players Dataset:](https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset) 3

The dataset contains over 18,000 players,700 clubs and 160 countries. The data was scraped from the FIFA 23 Official Player Database and includes information on players' personal details and FIFA-generated ratings.

It is debated to be the best representation of real world players ratings. Each player is carefully monitored by thousands of football scouts and reviewers. The ratings are then carefully tested to ensure that they accurately reflect a player's abilities and are fair and balanced across all players. This rich dataset provides a valuable resource for predicting and analyzing the 2022 World Cup.

## [FIFA World Ranking 1992-2022:](https://www.kaggle.com/datasets/cashncarry/fifaworldranking) 4

This dataset provides a robust view of the official FIFA International World Rankings for the last 30 years. For the purpose of this project I used the latest ratings for 2022.

# **Implementation Process:**

## Loading dependencies

* + Loaded in the dependencies used across the project.
  + This included a Kaggle specific dependency [opendatasets](https://pypi.org/project/opendatasets/)5.

## Loading in the data:

* + Loaded in the FIFA 23 and FIFA rankings using the opendatasets python library using an account specific API key.
  + Read the data in using .read\_csv()

## Data Cleansing

* + Filtered both datasets using a custom list of the 32 teams participating in the world cup using the .isin() function.
  + Created a ‘Position’ column grouping players into a set of three position types; Attack, Midfield and Defense as only specific positions existed ie: Right-Winger (RW).
  + Removed the columns in both tables deemed not necessary.
  + Normalized player names using .normalize() to remove umlauts etc. This could be important if the project was expanded later and was joined to a new dataset on a player’s name.

## Creating a Normalized Weighted Player Rating

* + Joined both tables together using .merge().
  + Split the joined tables by Positions using a for loop. This enables the creation a normalized rating for each position so positions like Attacker which have higher rated players on average are represented correctly.
  + Created a weighted rating for a more balanced and accurate representation of a player in the World Cup using the below:
    1. 60% - FIFA Overall
    2. 30% - FIFA Potential
    3. 10% - Value in Euro
    4. 10% - Wages in Euro
    5. 10% - National team’s world ranking
  + Normalized this rating between 70 and 95 (EA uses 0-100). This gave lesser teams a better advantage in the World Cup.

The resulting rating gave a strong left skew towards 70 rated players but still recognized the best players in the world.

### Chart, line chart Description automatically generatedFig. 1: Weighted Normalized Rating vs FIFA: Overall and Potential

Here we can derive that including a player’s Financial Wages and Player Value is correlated with their FIFA 23 Overall and Potential.

## Play\_game() function

* + This function plays to Nationalities against each other and predicts a winner.
  + Prepared the data for this by splitting each Nationality and Position into a dataframe using a nested for loop i.e. Brazil Attack, Midfield, Defense.
  + Created a mean rating per Nationality and Position.
  + Calculating a winner:
    1. Introduced random() and .norm() (normal distribution) to calculate the probability of a win.
    2. The Defense from Team 1 is compared to the Attack of Team 2, whichever is larger wins the round
    3. The Midfield from Team 1 is compared to the Midfield of Team 2, whichever is larger wins the round
    4. The Attack from Team 1 is compared to the Defense of Team 2, whichever is larger wins the round
    5. The team with more rounds wins the tie.
  + When a team wins the tie their team rating is increased by 1, when they lose it is decreased by 1.

## Simulating the 2022 World Cup

* + Recreated the group stage matches that are scheduled to play during the World Cup splitting the teams by their assign groups (Group A-H).
  + Took the winners from each group and seeded them into the Round of 16 and simulated the matches.
    1. From the round of 16 onwards all matches are Knockout games. To simulate this if two teams drew I created a play\_knockout\_game() function that would use random.choice() and select a team to ‘win on penalties’
  + Continued the process through Quarter Finals, Semi Finals and the Final until a winner was declared.

# **Results:**

### Fig. 2: Correlation between Player Value and their Weighted Normalized Rating

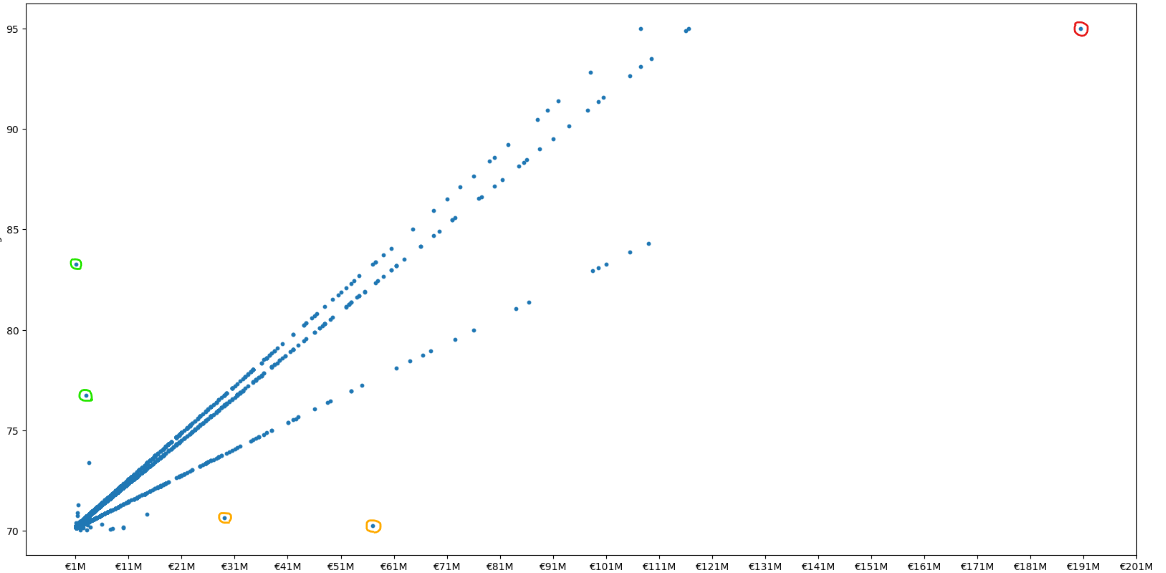


Figure 2 shows a strong positive correlation between the normalized rating on the y axis and Player Value on the x axis.

There are three main outliers in this correlation:

* Kylian Mbappé (red) – this wonderkid is said to be the next Messi/Ronaldo, he is both very highly valued in the real world and highly rated in the normalized rating.
* Overvalued players (orange) – These players under the line of best fit could be said to be overvalued as they have a high value but a low rating.
* Undervalued players (green) - These players have low values but a high rating. This an ideal financial investment for a club.

### Fig. 3: Pairwise plot of players values against their different ratings

Above we can see very strong positive correlations between a player Wage and Value to the three player rating stats used in this dataset. This points out the obvious, better players are worth and paid more than lesser players.

### Fig. 4: Graphic representation of the simulation of the 2022 World Cup created in Photoshop.

As seen in the Photoshop graphic above we can see the final output of the 2022 World Cup Simulation.

**Brazil win the Final** against Portugal.

At the time of writing this conclusion the Quarter Finals have just been completed.

* The function simulation predicted 10 of 16 of the Round of 16 results correct. Large favorites like Germany and Belgium where knocked out in the Group Stages which was unexpected.
* The function correctly predicted 5 of 8 of the Quarter Finalists.

### Fig. 5: Distribution of Positions in the dataset

Chart, pie chart

Description automatically generated

In the figure above we can see the distribution of positions across the dataset. This column was data mined using players specific positions. We can see in the graph that there are 10% more midfielders in the dataset then the other positions.

### Chart, scatter chart Description automatically generatedFig. 6: Pairwise relationships of the Top 10 Attackers across five key player stats

Above we can see a pairwise plot of the Top 10 attackers across multiple player stats.

The most interesting comparison in the graph is Robert Lewandowski (grey) and Vinicius Jr. (orange).

In the top right pairwise plot we can see that Lewandowski has a very high Finishing and Strength rating whereas Vinicius has a much lower Strength and Finishing Rating.

Whereas in the plot in the center to the right we see a large difference again where Vinicius has a very high Sprint Speed and Stamina yet Lewandowski is much lower in this regard.

This is a perfect example of two very high rated attacking players that have different builds and styles of play.

### Fig. 7: Pairwise relationships of all players depending on their position

Shape

Description automatically generated

There are clear trends in the graph above.

It can be surmised that:

* Defenders have the best Strength stats
* Midfielders have the best Stamina and Short Passing stats
* Attackers have much better Finishing and Sprint Speed stats

### Chart, calendar Description automatically generatedFig. 8: Heatmap of the average player per country

After a quick analysis above, evident trends can be seen:

* Brazil, France and the Netherlands have the best average stats
* Spain and France have the best Short Passing while Australia and Brazil have the best Stamina
* Netherlands and Cameroon are the Strongest teams.
* Serbia stand out from all other nations to have the lowest Sprint Speed on average
* Canada and Qatar have the worst Finishing
* Qatar the hosts of the 2022 World Cup have the worst overall stats. Host nations are given automatic qualification to the tournament

### Fig. 9: New Team Rankings vs FIFA Ranking Difference

Chart

Description automatically generated

Analyzing the above we can see:

* Serbia and Cameroon have the biggest gain in Team Rankings. This is due to the teams in their group being much weaker than others.
* Teams that got very far in the tournament like England and Argentina have dropped the most. This is due to the calculated ‘form’ penalizing them for playing more games in the tournament. Adjusting ‘loss’ form for knockout stages could be an improvement to better this.

# **Insights:**

* Data analytics can be used to create accurate and comprehensive financial analysis in football by collecting and analyzing data on a wide range of financial factors. By using advanced analytical techniques, such as machine learning and predictive modeling, data analysts can uncover valuable insights and trends that can help teams and organizations make more informed financial decisions – see [figure 6.](#_Fig._6:_Pairwise)
* Data mining and analyzing can easily identify undervalued for purchase. Desired player strengths can be identified quickly to fill gaps in a team. This information can be valuable for scouting purposes, as it can help teams prioritize which players to target in transfers or other acquisitions, see [figure 2](#_Fig._2:_Correlation).
* Predicting the outcome of a game between two teams is very difficult. Even with very extensive predictive models like ones used by bookmakers there is unpredictably that cannot be identified. This flaw can be seen in this projects results in [figure 4](#_Fig._4:_Graphic).
* EA’s ratings are not always perfect. As seen in [figure 2](#_Fig._2:_Correlation) there are still overvalued and undervalued players within their ranks. These ratings continue to improve with the increasing accessibly of data in football. For example, in the 2022 World Cup, Adidas developed a special ball that contains IMU sensors that collect data points on the ball and use it for their semi-automated video assisted referee system.6
* Strong correlations can be seen for players’ stats in different positions as seen in [figure 7](#_Fig._7:_Pairwise). With further analysis and comparison of more player attributes, clusters of different types of players could be identified in each position. For example, a holding midfielder or an attacking midfielder.

# **Machine Learning and Future Development:**

Future developments in this project could include:

* Classification models and an expanded dataset the predictive model could be expanded for more accurate results.
* A combination of linear and non-linear regression dependent on whether the players stats are obtained playing for their club or national team.
* Automating the player scouting process by training machine learning models to identify and evaluate potential transfer targets based on their attributes and performance data.
* Expanding the project to different sports like the NBA and NHL.

# **References:**

1. <https://www.masterscompare.co.uk/masters-courses/performance-analysis-in-football/156838/>
2. <https://www.vg247.com/2016/09/27/how­ea­calculates­fifa­17­player­ratings>
3. [EA Sports FIFA 23 Players Dataset:](https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset)
4. [FIFA World Ranking 1992-2022:](https://www.kaggle.com/datasets/cashncarry/fifaworldranking)
5. [opendatasets](https://pypi.org/project/opendatasets/)
6. [Introducing ‘Al Hilm’ - the official finals match ball (fifa.com)](https://www.fifa.com/fifaplus/en/articles/official-match-ball-world-cup-finals-al-hilm-adidas)