Soccer Performance & Rating Data Exploration

Emre Guvenilir & Chadi Bsila

Datasets

Utilizing two different datasets

- EA FC 24 Player Database
  - Dataset contains players from FIFA 15 to EA FC 24
  - Contains metrics such as overall rating, pace, shooting, passing used as our target values
  - Subjective, data-driven ratings given by EA FC data analyst team
  - Few options available for complete season performance ratings
- 2022-23 Football Player Stats Database
  - Dataset for big 5 leagues (Premier League, La Liga, Serie A, Bundesliga, and Ligue 1)
  - Contains both basic and advanced metrics and player information for our features
  - We might not proceed with this dataset and either create our own through a FotMob API
    - Dataset contains information for about half the season, but we can still do per 90 stats

Goals of the Project

So this is going to be a little complicated... but just bear with us

- Develop models for players to determine what their EA FC rating will be in the next game
  - Forwards
  - Midfielders
  - Defenders
  - Goalkeepers
- EA FC 24 ratings are based on a conglomerate of factors from the previous season, and heavily rely on team and individual player performance
- So, given we know the ratings of the previous game, we want to train a model on the data from that season to see if we can determine what rating they will get in EA FC 24, which we have their *actual* EA FC 24 rating

Features

Feature choice will differ depending on the position group or model. For example:

Forwards: Goals, assists, shots on target, press completion %, fouls drawn, etc

Midfielders: Passes that enter final 3rd, goals, assists, pass completion %, etc

Defenders: Goals, assists, aerial duels won, blocked shots/passes,, fouls made, etc

Goalkeepers: Saves, passes made under pressure, goals conceded, penalties saved, etc

All models will contain features such as age, matches played, starts, minutes, since all these stats are often basic indicators of performance when used in conjunction with other chosen metrics

These are by no means an exhaustive list, but can help give an idea of how different positions need to be evaluated differently

Data Processing

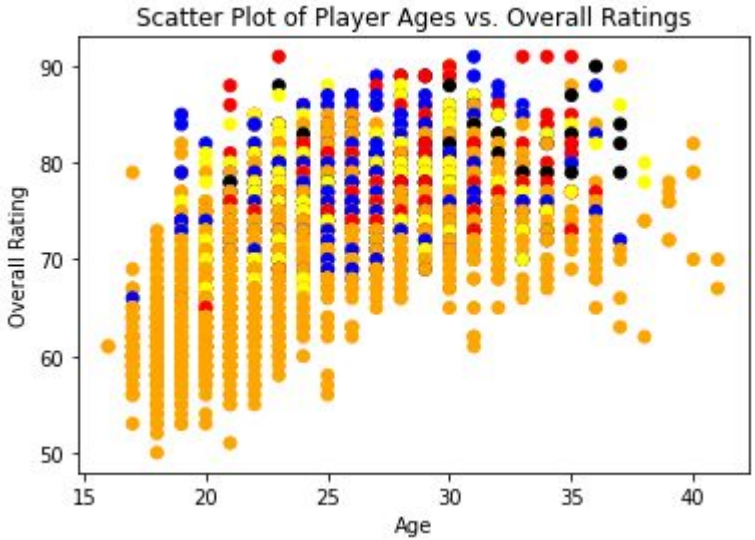
Filtered EA FC 24 dataset to only players in the big 5 leagues, and then created 2 separate dataframes for FIFA 23 and EA FC 24 players, respectively

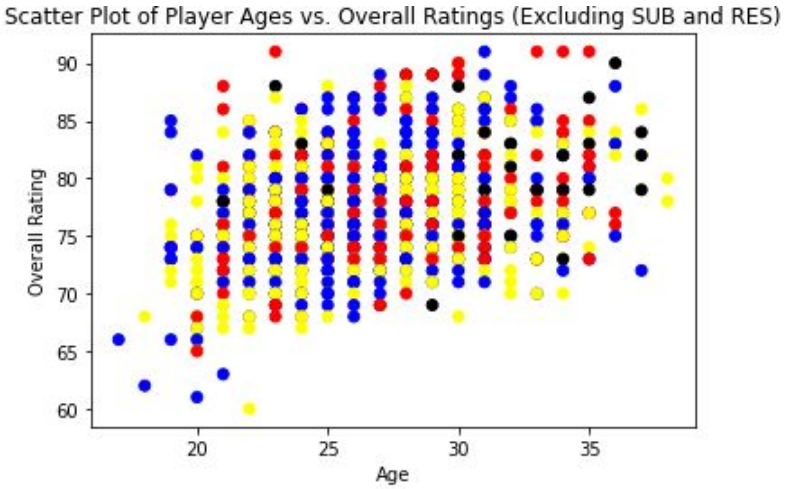
We used a name matching process utilizing process.extractOne() and a for loop to match our two datasets and add rating as a column, since ultimately only players who we had real life stats for could be considered

- We were not able to directly match strings as names with accents or other characters were corrupted in the player stats dataset
- The EA FC 24 dataset contained a players full name, meaning we used a ‘best match’ system along with checking the team and age were the same
  - This ensures players with the same name could not be mixed
  - Dembele, Lee, Park, Traore are common names and it is not unusual for players to have the same name

We also thought EA FC’s positions were clearer and more accurate, so we also added that as a column which we will eventually split data on further for each of the models

Trends

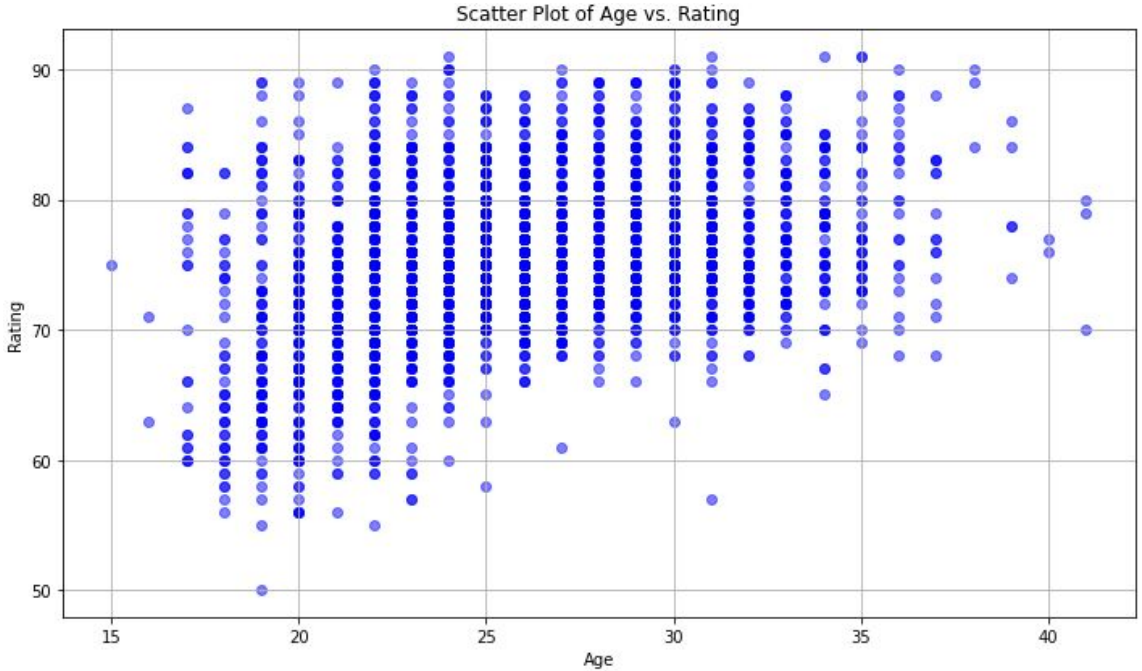




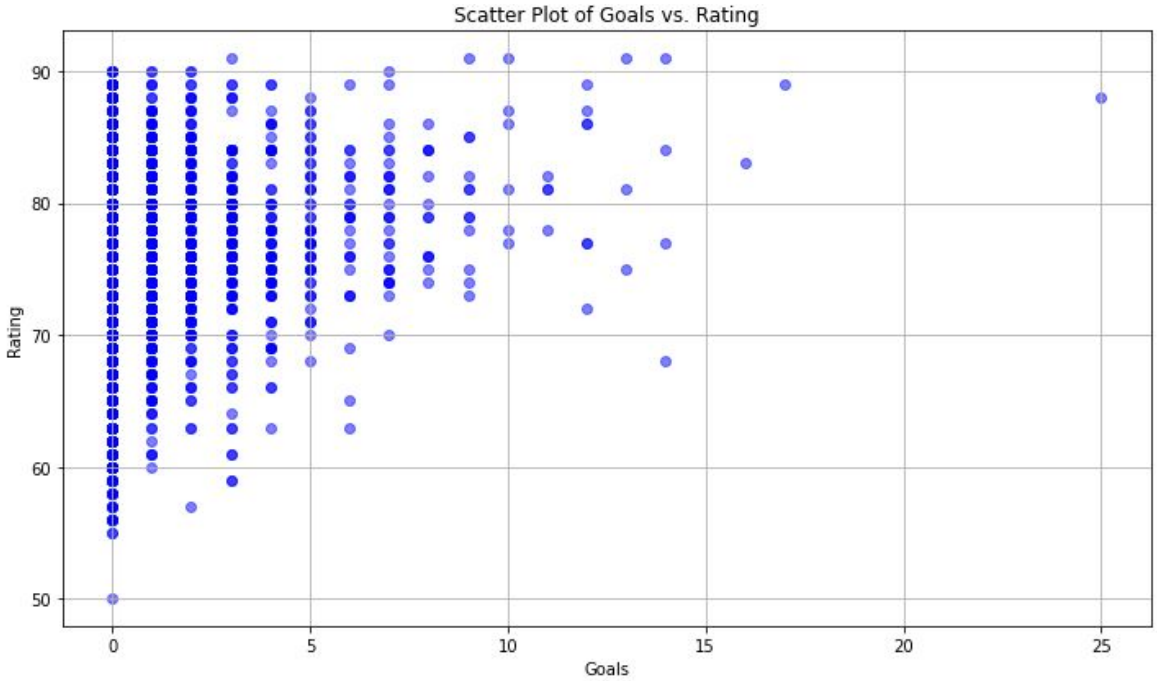
Red: Forwards Blue: Midfielders Yellow: Defenders Black: Goalkeepers Orange: Substitute/Reserve

*Positions of players were determined from EA FC 24 dataset, not the player stats

Trends (cont.)


Scatter Plot of Age vs. Rating
Rating
Age
90
80
70
60
50
15
20
25
30
35
40

Trends (cont.)


Scatter Plot of Goals vs. Rating
Rating
Goals
90
80
70
60
50
0
5
10
15
20
25

Model Types

- Linear Regression (with and without NN)
- Polynomial Regression
- Ensemble Methods to compare performance of different models as our project is highly dependent on which features you choose.
- SHAP values to compare performance of initials models on selections of different features.
- Clustering players based on similarity.

Ethical Considerations

Biases of rating models:

- Younger players tend to be rated lower as they are new to the professional game, even if they have a better year than older players
- Players from countries outside of the top leagues can often be under scouted and underrated
- Previous studies claim FIFA ratings have shown to perpetuate stereotypes between black and white athletes, although the game developers dispute this finding
- Rating models tend to favor attacking players, especially with how highly goals and assists are weighted

Application considerations

- Ratings can affect the public perception of a player as FIFA is a widely popular game
- Statistical analysis is often used when players are negotiating new contracts
    - Our model must be fair in evaluating the performance of each position type