Jason, 1B, Feature analysis and Simple Classification Model.

Analysis:

Before starting on the Classification Model, we first had to do some analysis into the features in the dataset. Below are some graphs I used for analysis and their interpretations.

Afbeelding met tekst, schermopname, diagram, Lettertype

Automatisch gegenereerde beschrijving

Here we visualize the top 5 teams based on Average Market Value. We can see here that the average Market Value of Liverpool is very high compared to the rest, followed is West Ham United with a Market Value which seems to be around 60% of Liverpool's.

Afbeelding met tekst, schermopname, Perceel, Lettertype

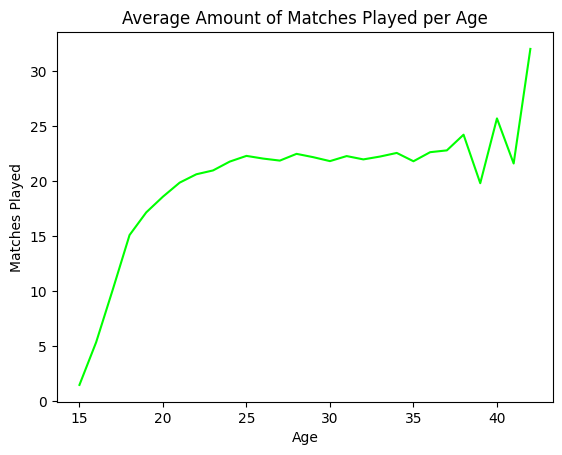
Automatisch gegenereerde beschrijving

Next we visualized the distribution of Positions between players. Here we can see again that CF is the most popular position by far.

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

I chose to only visualize the top 5, just so the graph will stay readable and won't overflow you with information. In this visual we can see that the country on first place is Italy, the amount of players from Italy heavily outweigh the other countries.



we can see that the younger a player is, the least matches they play, but this changes after around the age of 23: from this point onwards the line seems to be staying around the same value. There seems to be random peaks of data after the age of 35, but this is due to the lack of players in that age group in the dataset.

Afbeelding met tekst, schermopname, lijn, diagram

Automatisch gegenereerde beschrijving

How do duels won and aerial duels won vary by position? Looking at the visual, we can keep the conclusion made earlier: the more defensive positions have the higher average amount of duels won and both datapoints have about the same distribution.

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

I can see that there is a really slight correlation to the two variables. There has to be a different variable which also has impact on this correlation, but it needs some extra future research.

Machine Learning:

First we started making a basic model without changing (preprocessing) the data too much, which gave the following results per type of model:

Linear Regression:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Logistic Regression:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Random Forest:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

Gradient Boosting Trees:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

To improve these models I applied Feature and Correlation Analysis per each position:

­Goalkeeper:

Afbeelding met schermopname, plein, tekst, patroon

Automatisch gegenereerde beschrijving

Attackers:

Afbeelding met schermopname, tekst, plein, Kleurrijkheid

Automatisch gegenereerde beschrijving

Defenders and MidFielders (combination)

Afbeelding met tekst, schermopname, plein, software

Automatisch gegenereerde beschrijving

Afterwards, I ended up scaling the data to make sure that they can be used along each other better, which often results into higher performances:

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

And afterwards, I applied parameter selection for certain models, below I will only show one example, since I did this using 2 methods for 2 models for each position.

Afbeelding met tekst, schermopname, software, Multimediasoftware

Automatisch gegenereerde beschrijving

After this, I trained the new models with updated data and better parameters, which ended up with the following performances: (Note: These performances were for goalkeepers, I had a different model for every type of position.)

Afbeelding met tekst, schermopname, software

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, software

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, software

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermopname, software

Automatisch gegenereerde beschrijving

These are the final performances for goalkeepers. We print out multiple types of scores, so we can decide which model is best fitted for the question based on multiple factors and performances.