DATA 1030 Midterm Report

Kaishuo Zhang

Introduction

In this project, I used the dataset named “All Time Premier League Player Statistics”. I am going to explain how I predict a player’s best position using the statistics of their career in the English Premier League. The target value is the position of a given player, which is defined as forward, midfielder, and defender. Hence, it is a classification problem. This result could be used when deciding what position a player should be playing based on his performance on the pitch. For instance, if a defender is producing stats like a midfielder, it would be a good indicator that he will be a better fit in the midfield. This problem is important because when deciding a player’s best position on the field, we can only look at his previous performance and many players found success after the transition to a new position at some point in their career. This data set was found on Kaggle, the author obtained the data from the official website of the premier league. There is a known issue that some of the players have incorrect goals per match stat and it will be discussed in the EDA section. Thanks to the best data collection technology of the premier league staff, the author was able to get 502 players(excluding Goalkeepers) and more than 50 different stats.

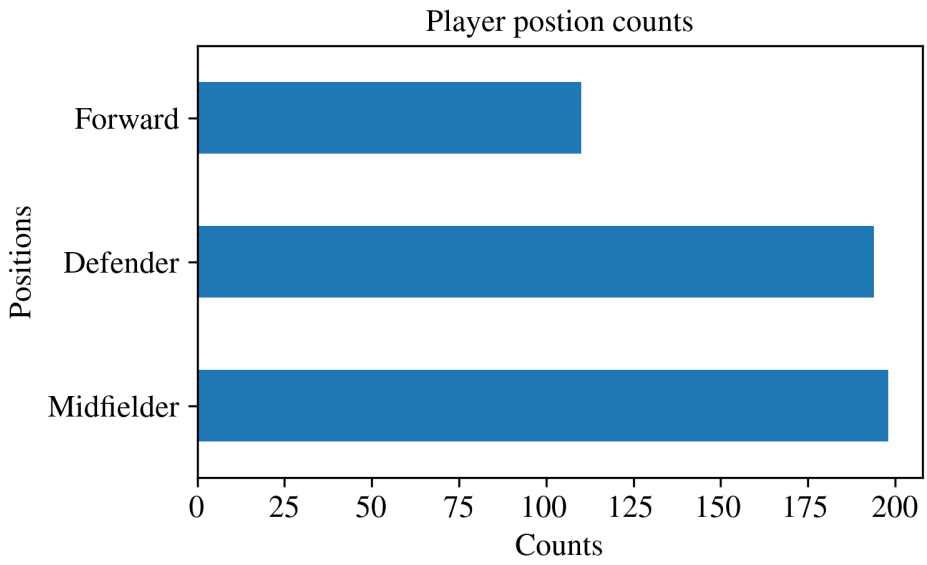
The dataset contains game stats associated with each player in different areas on the pitch. For example, shot attempts, goals, interceptions, and passes completed are all included. Some other interesting features are also recorded in this dataset, such as the player’s nationality and jersey number. It is fascinating to see if these also have impacts on deciding what position the player is best suited for. In this data set, nationality, jersey number, and club are categorical features and all other features are continuous.

Even Though this dataset is found on Kaggle, there has been no similar research in this manner. Most people use this dataset to visualize the players’ performance as a group in the league and also individual players’ performance while trying to show how much better or worse a player is compared to his peers.

Exploratory Data Analysis

With the incorrect data mentioned in the last section, I found that all the impacted players had 0 goals in the league and for some reason, their appearances are used in that column. So I filtered all the players that have more than 1 goal per game in the dataset and set the value equal to 0 since the most goals per game stat in the history of the premier league is less than 1.

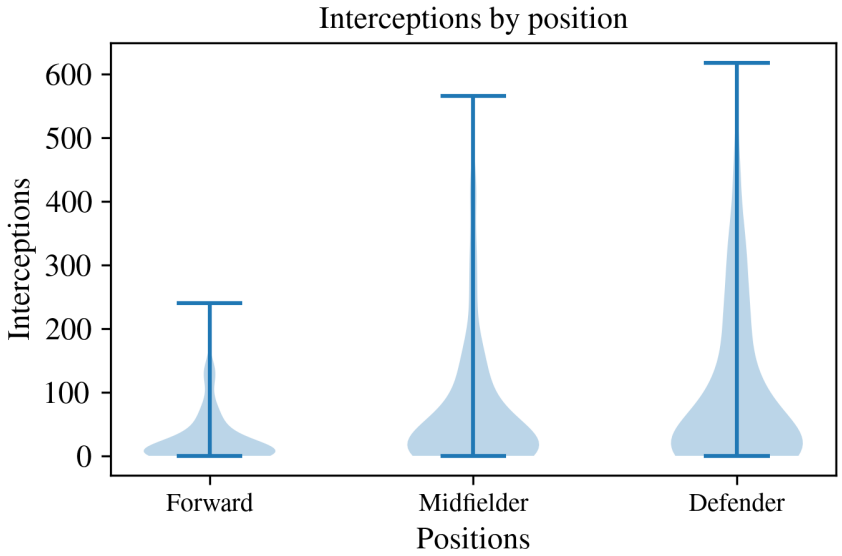
In terms of my target variable, there are three categories, ‘Forward’, ‘Defender’, and ‘Midfielder’. As shown in the graph below, the dataset is imbalanced as there are significantly fewer Forward players. This should be considered when splitting the data to avoid further problems.



This graph shows the distribution of players are in each position

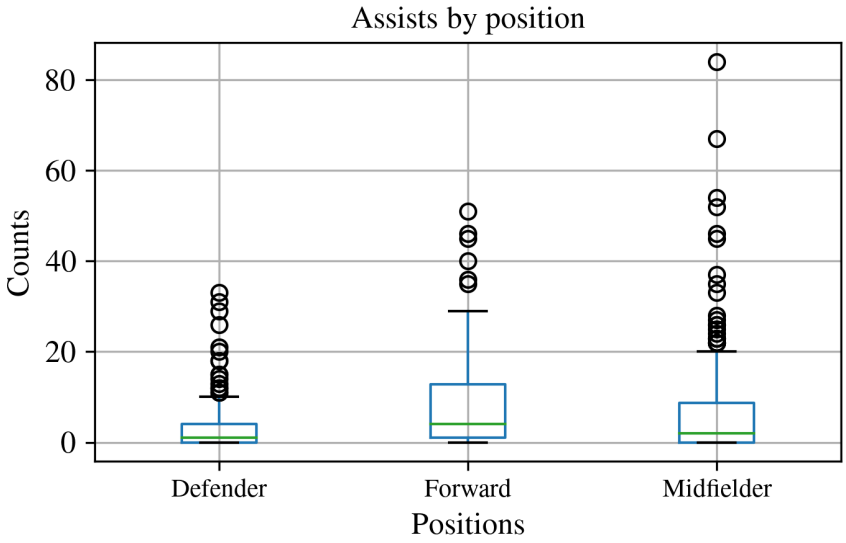
Some other interesting features are interceptions and assists because they are closely related to a player's position. I used the violin plot and the box plot below to show each feature's correlation to the target variable and the results are very interesting.

The expectation is that a player playing in a deeper position should have more interceptions than a player playing a more advanced role. As shown in the graph below, defenders have the most interceptions among all three positions while forwards have the fewest. This perfectly fits our expectations and it indicates that this feature could play a big role in the machine learning pipeline.



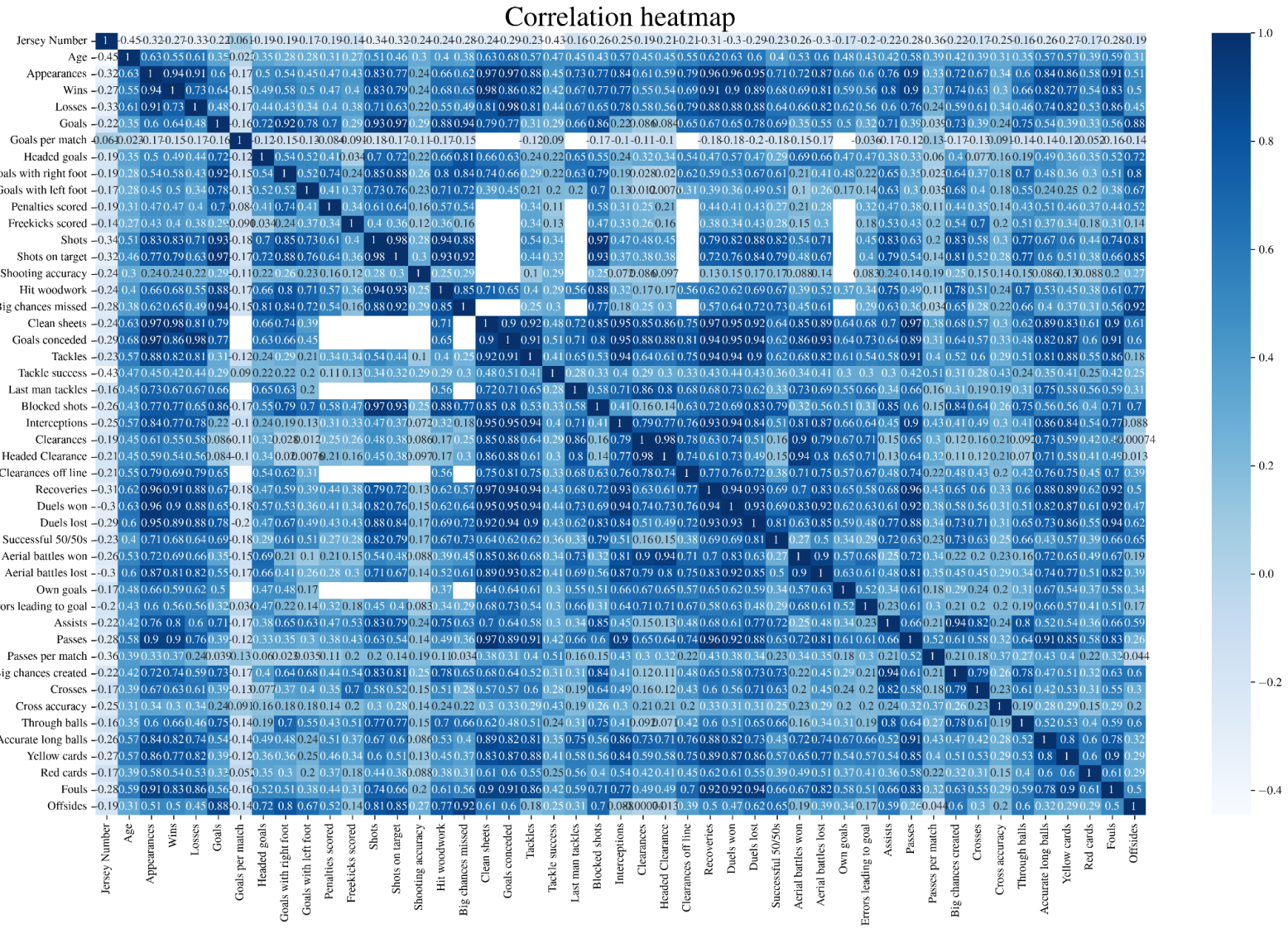
This graph shows the correlation between interceptions and positions

On the other hand, assists should be the opposite of interceptions where a player playing up the field should have more assists than a player playing in the backline. It would be easier for them to make an assist since they play closer to the opponent's goal. As you can see in the graph, defenders have the fewest assists of the three positions. While forwards and midfielders have approximately the same median, midfielders have a heavier tail because of outliers. This does make sense because most forwards love to shoot the ball instead of passing it and some midfielders have the pass-first mentality which helps some of them to have the most assists in this data set.



This graph shows the correlation between assists and positions

I also made a heatmap to see if there are any high correlation features in my data set. And as a result, the highest correlations are between headed clearance and clearance, and between shot and shot on target. I do not believe these correlations are high enough for me to drop one of them before the machine learning pipeline so I would just leave them there and make a decision later.



This graph shows the correlation between features

Data preprocessing

After doing some EDA, it is very clear that my data is IID and has no group structure. And as a result, I chose to divide my data into a train set, a validation set, and a test set with the size 301, 100, and 101, respectively. In terms of the dividing method, I used stratified K Fold since I want to make sure each class is represented with a fair amount in each of the groups given the fact that my data is imbalanced. This eliminates the possibility that all forward players fall into test and validation sets which will affect my model.

When encoding the features, I used OneHotEncoder on all of my categorical features, including “Jersey Number”, “Club”, and “Nationality”. I used MaxMinEncoder on “Age” because it has a very clear bound. For all the other continuous features, I used StandardScaler. Finally, There will be 176 features in the data set.

In terms of missing values, I imputed 0 for some of the features when it makes sense to do so. For example, many players have “penalty scored” as missing and that is most likely because they have never taken a penalty in their entire career. There are a couple of other features that have the same characteristics and I did the same for those. For all the other features, I left them as missing and will deal with them later in the pipeline.

References

https://www.kaggle.com/datasets/rishikeshkanabar/premier-league-player-statistics-updated-daily