OCSC Intern Project

Michael Lucio

For this project, I decided to go with **Option 1** - Data Engineer. The problem statement for this project was to construct a CSV containing a timestamped list of Fullback Overlaps, ranked in order of effectiveness, from your last match (ORL - CHI : 2022-4-9). I broke this project up into three distinct parts: data processing, identification of the overlapping runs, and evaluation of these runs. Let's take a closer look at each part.

1. Data Processing

I decided to do this project in Python, so I began with importing the tracking data and metadata into a Jupyter notebook. I loaded the metadata and kept it as a json, but ended up separating the *homePlayers* and *awayPlayers* dictionaries into separate pandas Dataframes as they might be easier to work with in the future. I loaded the tracking data directly into its own pandas Dataframe, but again ended up further filtering out the Orlando City and away team data into separate Dataframes.

In soccer, the teams switch sides during halftime and end up attacking in the opposite direction the second half of the game. To deal with this, it is important to inverse the spatial data, such as the coordinates of the players and balls, for the second period of the match. With this change we ensure the teams are always on the same side.

The last step to the data processing was to identify the Fullbacks, the players we are investigating in this problem, and filter the data frame to only show their tracking data. It was easy to find the starting fullbacks because they were listed in the metadata, but were there any substitutions? To investigate this further I built a function to calculate total play time and found that Ruan (#2, RB) only played 75 minutes and was replaced by K. Smith (#24). Our LB is Joao (#4).

2. Identifying Overlapping Runs

Using the definition of a fullback overlap given in the prompt, I built a function to identify the initial and final frames of each overlap a player makes. The function takes in a player's dataframe as the parameter, and it iterates through it as it checks for the right criteria to be labeled an "overlap." The criteria I set for an overlap are as follows:

a. The player's speed must be at least 5.5m/s for at least 3 frames.
b. The player must initially be behind the ball, and run forwards.
    i. Player's Initial X > Player's Final X
    ii. Player's Initial X > Ball's Initial X
c. The player must run towards the exterior of the pitch and around the ball.
    i. If Player's Initial Y is positive, then the Player's Final Y > Initial Y
    ii. If Player's Initial Y is negative, then the Player's Final Y < Initial Y

3. Evaluating Player Runs

In order to evaluate each run we had to calculate two things: Pitch Control (PC) and Expected Possession Value (EPV). Pitch control is the probability that a team will gain possession if the ball is moved to that location on the field. Expected Possession Value calculates the likelihood of scoring with possession of the ball on any given location on the pitch. When we multiply the two together, PC * EPV, we can get a good idea at what is going on behind the scenes and add value of off the ball movement. In this case, we will calculate PC*EPV for the final frame and compare it with PC*EPV of the initial frame of each run to get our evaluation.

Final Thoughts & Conclusion

This was a very fun project as it made me connect my data science skills and apply them to the world of soccer analytics and tracking data. I had studied a lot to understand analysis concepts you can get from tracking data such as Pitch Control and Expected Possession Value. I can say proudly that I learned a lot throughout this project and struggled a lot at times, but this is what I am passionate about.