Unlocking Potential:
Identifying Data-Driven Ways for an Aspiring Player to Improve Their Perceived Value and Potential

Michael Govaerts

Dataset(s)

Soccer Dataset

(European Soccer Database from Kaggle – Author: Hugo Mathien)

https://www.kaggle.com/hugomathien/soccer

Motivation

Pre-professional athletes are often competing against a large pool of their peers for a very limited number of professional positions. In many cases, these players have worked for more than a decade to hone their sport-specific abilities and get noticed by teams that could take them on. As the probability of success in this regard is dishearteningly low, any evidence-based guidance in improving a player's odds for selection would be extremely valuable. Instead of operating off hunches, superstition and perceived trends; data-driven insights to help a player improve their perceived potential would be incredibly valuable for the player and what is often their dedicated network of supporters.

To help supply these insights, data science principles can be used to analyze data from current players and their potential rating to help determine if any particular player attributes are notably related. Armed with the European Soccer Database and its catalog of more than 10,000 players and their statistics, we hope our analyses can help identify any relationships that may help these players in their professional pursuits.

3

Research Question(s)

When comparing professional player attributes to a player's rating and perceived potential, do any meaningful trends emerge when focusing on a large sample of European professional soccer players?

Can these conclusions be extended to the pre-professional community to help them select more valuable areas of improvement to enhance their rating and perceived potential in the hiring process?

4

Findings (1/6)

After cleaning the data set, correlations between the various player attributes and the two overall player quality measures, namely a player's overall rating and perceived potential, were calculated.

We created two graphs to help make the strongest correlations easier to identify across the numerous player attributes.   One graph was for attribute correlations with player potential, and the other was for attribute correlations with overall player rating. The strongest correlations visible in the the graphs and obtained from the data are listed below.

Player Potential Correlations

```
reactions: 0.580991
ball_control: 0.401803
vision: 0.379278
short_passing: 0.382538
long_passing: 0.343133
positioning: 0.326898
penalties: 0.315207
shot_power: 0.325459
```
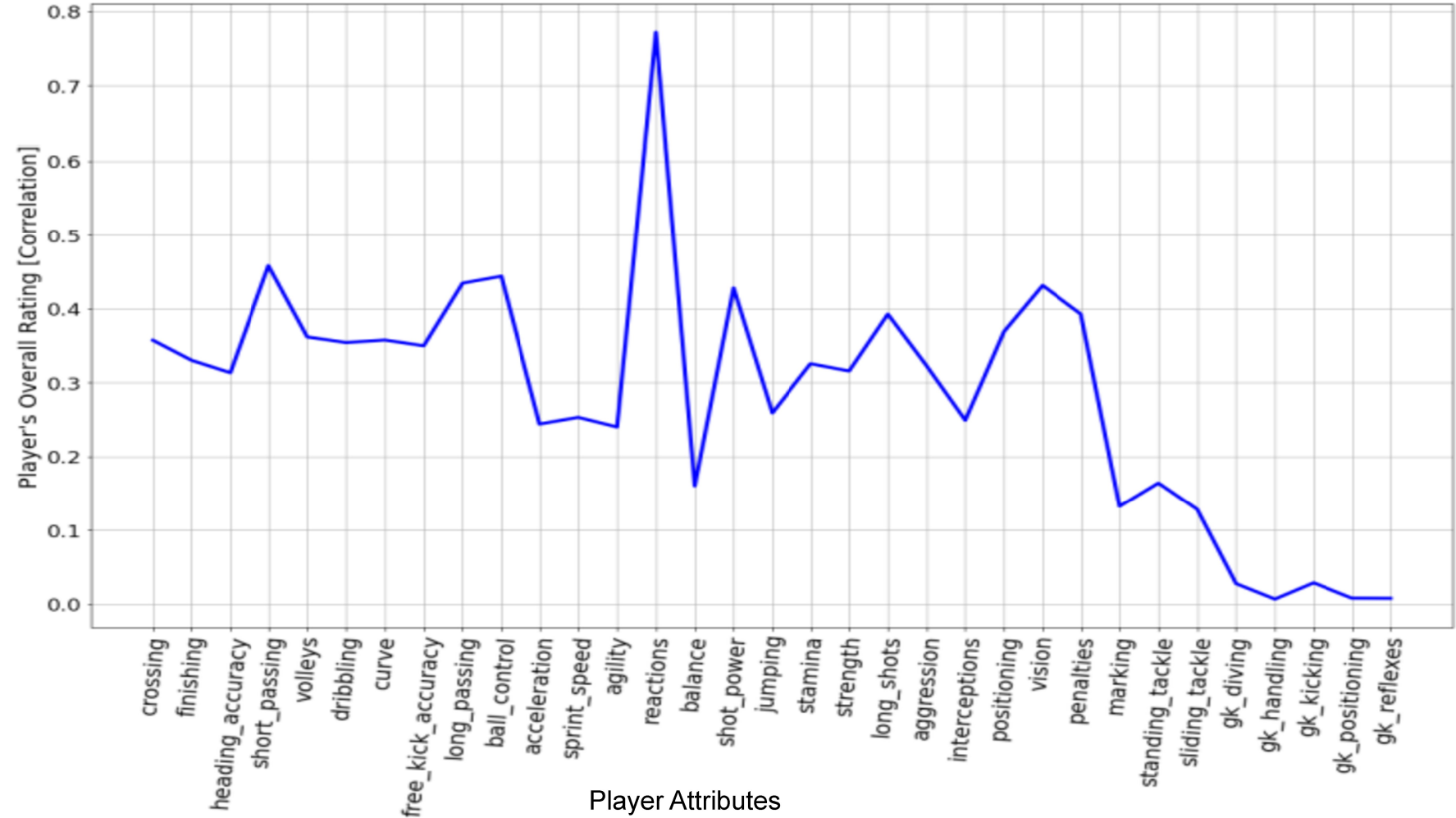
Player's Overall Rating Correlations

```
reactions: 0.771856
short_passing: 0.458243
ball_control: 0.443991
long_passing: 0.434525
vision: 0.431493
shot_power: 0.428053
```

Player Potential Correlations
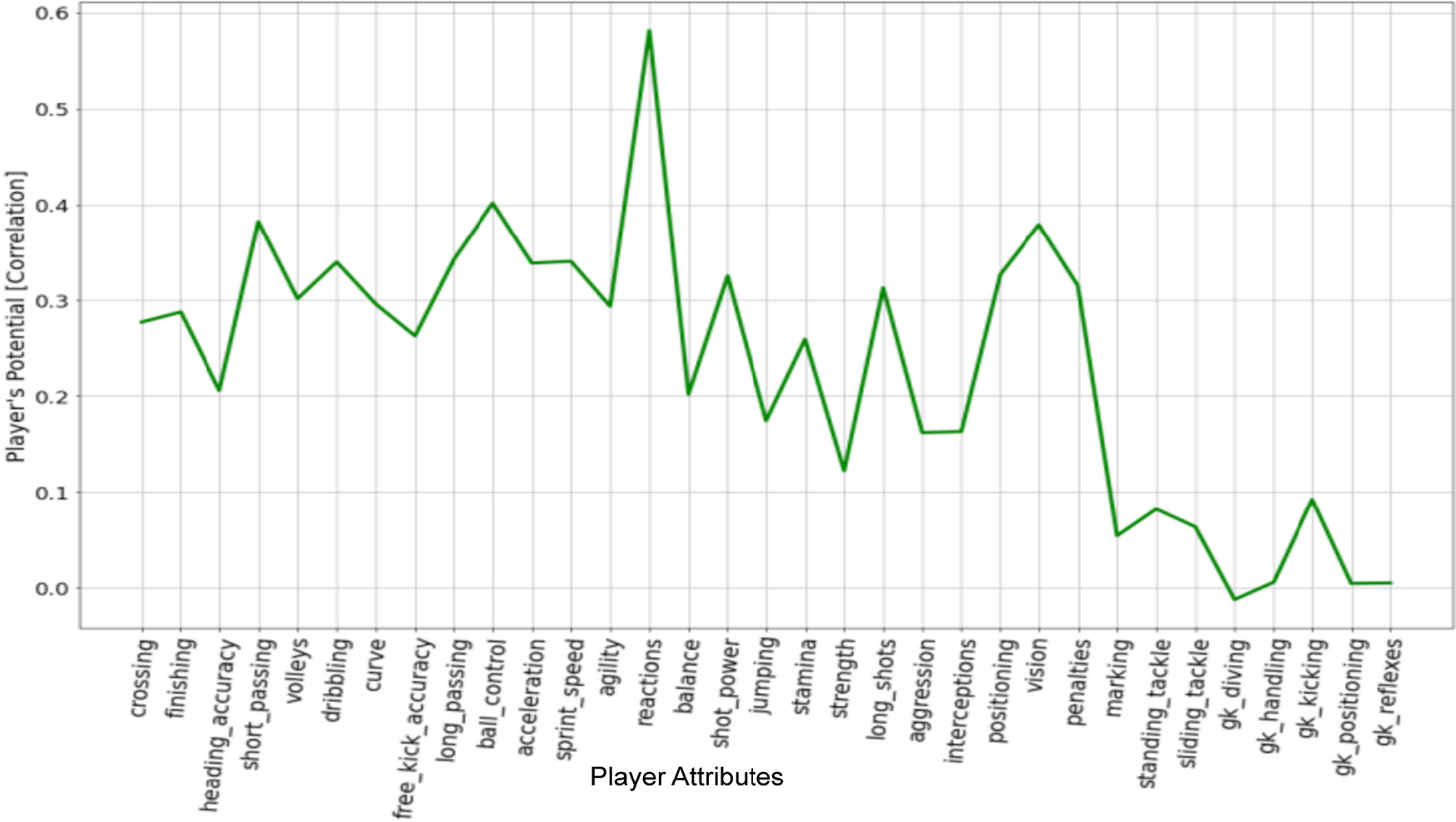
Player's Overall Rating Correlations

Through visualizing these two graphs, we identified that the strongest correlations between player attributes and overall player performance measures occurred in the attributes for 'short_passing', 'long_passing', 'ball_control', 'reactions', 'shot_power', 'positioning', 'vision', and 'penalties'.

The reactions attribute was correlated with player potential and overall player rating at 0.581 and 0.772, respectively. The reactions attribute is defined as follows:

Reactions: "Good reactions will allow you to latch onto loose balls & rebounds quicker than other players." [as defined by the sofifa.com website referenced by the Kaggle European Soccer Database page]

8

Player's Overall Rating Correlations

Player Potential Correlations

```
reactions: 0.771856
short_passing: 0.458243
ball_control: 0.443991
long_passing: 0.434525
vision: 0.431493
shot_power: 0.428053
```

```
reactions: 0.580991
ball_control: 0.401803
vision: 0.379278
short_passing: 0.382538
long_passing: 0.343133
positioning: 0.326898
penalties: 0.315207
shot_power: 0.325459
```

This reactions attribute was much more highly correlated than the next strongest correlated attribute for each performance value, namely short_passing (correlations of 0.458 with player rating and 0.382 with player potential) and ball_control (correlations of 0.444 with player rating and 0.402 with player potential).

Our correlation heatmap helps further visualize these relationships and their relative strengths.

9

Correlation Heatmap

Critical to our research questions, we can see the reactions attribute is notably more correlated with player potential and a player's overall rating than the other attributes. Specifically, the reactions attribute has an even stronger correlation with overall player rating, with a correlation of 0.77. Furthermore, this 0.77 correlation is much stronger than the second strongest overall player rating correlation with the short_passing attribute at 0.46, identifying a clear trend in these attributes.

Thus, our analysis has clearly identified that some player attributes are more strongly correlated with player potential and overall player rating than others. Therefore, we identified meaningful trends in the attributes' associations with overall player performance measures, which answers our first research question concerning the presence of any such trends.

Additionally, our analysis also suggests that our conclusions can be extended to the pre-professional community, addressing our second research question. Using our data, pre-professional athletes and their advocates can focus on improving player attributes most strongly associated with a professional player's overall quality and potential. By focusing on enhancing these more strongly correlated attributes, the pre-professional players can hopefully improve their chances of drawing professional soccer's attention and subsequently entering the professional soccer community. Based on the notable strength of the "reactions" attribute in this project, pre-professional players would likely be wise to improve their own abilities in the "reactions" attribute to enhance their own perceived overall player rating and potential in the eyes of professional soccer.

However, to draw more comprehensive and definitive conclusions, we would likely need to utilize more comprehensive statistical methods and analyses to effectively carry out sensitive statistical tasks like removing any confounders, identifying any issues with the data, and revealing more meaningful findings. Thus, further research with this data set and with additional data is likely needed to shed more light on our research questions and their implications for the professional and pre-professional soccer communities.

Acknowledgments

The author would like to thank the Kaggle for hosting the European Soccer Database data set used in this analysis. Also, we would like to thank Hugo Mathien for gathering the data, creating the data set, and sharing it publicly on Kaggle.

Also, we are incredibly grateful to the scientists, especially data scientists, who are kind enough to share their knowledge and expertise with the rest of us. Reviewing various data scientists' publicly available work and observing their online discussions significantly helped inform this project's direction.

Thank you, as well, to the instructors and participants in this course for allowing the author to become more proficient in data science.

References

European Soccer Database from Kaggle [Hugo Mathien]
https://www.kaggle.com/hugomathien/soccer

EdX / UC San Diego MicroMasters in Data Science: Python for Data Science
https://www.edx.org/course/subject/data-science