Soccer Data Analysis

Tufei Cai, William Fitzhugh

Author contributions

1. William worked on tidying the dataset, prepared the data description, create some data visualization, and wrote the background information.

2. Tufei worked on tidying the dataset, created some data visualization, finish exploratory analysis.

Abstract

Prepare an abstract *after* you've written the entire report. The abstract should be 4-6 sentences summarizing the report contents. Typically:

- the first 1-2 sentences introduce and motivate the topic;
- the next 1-2 sentences state the aims;
- the next 1-2 sentences state the findings.

Background

Soccer is the most popular game in the world. Played by billions of people, almost everyone has some knowledge of the game and their country's league. However, soccer is played differently throughout the world, with certain countries known for their flair and others known for their discipline. The goal of our project is to explore how these differences manifest themselves in data. The main data we will be working on is a collection of match results from eleven different European soccer leagues over a number of years (roughly 2010-2016). We will combine this data with information on player abilities, team play styles, and betting odds, to create insights into how these top leagues differ from one another. The infomation we create in this project could be used to aid the accuracy of a model used to predict soccer results.

Aims

The first approach will be are winning and scoring patterns the same across different European soccer leagues?(home advantage, goals scored, variability of top teams each year).

The second approach is the player and team attributes correlated with success according to FIFA soccer player data. By focusing on the physical characteristics of players who had a 90 or higher overall rating we have found out that the physical characteristic is not straight related to the win rate or player's overall rating. But it still has little correlation that can lead to the win rate or overall rating a little bit higher. By using linear regression as the tool to predict results we have found out that a player's height and weight play an important role in the soccer game. In other words, a player's physical characteristic is going to increase the win rate which means players with better physical data had a higher chance to win.

Materials and methods

The goal of this section is to describe your dataset(s) and sketch out your analysis.

Datasets

The match data is contains the home and away teams involved in the match, the goals scored by each team, and the result information. This data comes from a kaggle project (https://www.kaggle.com/datasets/hugomathien/soccer) which aimed to collect data on European club soccer. Data

was collected via web scraping, as all of this infomation is readily availbale online, but needed to be collected and stored. The relevant population is the eleven leagues for which matches are recorded, all of which are in Europe. We will focus on the biggest leagues, being England, Spain and Italy. The scope of inference is small, if anything, because this data is years old. If this were up to date data, then our finding would have some bearing on how these leagues will perform going forward, but given that this data is at least five years old, the leagues would have had enough time to change that our finding will likely have little to no bearing on European soccer today. Data that we will pair with the match infomation is the player and team attributes, which are taken from Fifa, the most popular sports video game in the world. Fifa is a soccer video game which maticulously tracks player abilities and team styles throughout the year. While these stats are obviously not a direct representaion of real life, they provide a good idea of how teams should play on paper. This data was also web scraped, and was found on the same kaggle page.

Data Structure

Dataset 1 :

The following data demonstrate the question of winning and scoring patterns the same across different European soccer leagues?(home advantage, goals scored, variability of top teams each year).

> **Table 1**: Variable descriptions and units for each variable in the dataset

Variable	Description	Units
match_api_id	Match id	Individual game
id	League ID	Individual ID for Players
League	League Name	NaN
Country	Country Name	NaN
season	Season of match	Soccer calender year (fall to summer)
stage	stage of season	Number game of season
home_team_goal	Goals scored by home team	Count(n)
away_team_goal	Goals scored by away team	Count(n)
home_team	Home team name	NaN
away_team	Away team name	NaN
winner	Winning team name	Nan
home_points	Points earned by home team	Win -> 3, Draw -> 1, Loss -> 0
away_points	Points earned by away team	Win -> 3, Draw -> 1, Loss -> 0

> **Table 2**: Example of relative abundance data.

match_api_id	id	League	Country	season	stage	home_team_goal	away_team_goal	home_team	away_team	winner	home_points	away_poin
492473	1	Belgium Jupiler League	Belgium	2008/2009	1	1	1	KRC Genl	Beerschot AC	Draw	1	1
492474	1	Belgium Jupiler League	Belgium	2008/2009	1	0	0	SV Zulte-Waregem	Sporting Lokeren	Draw	1	1
492475	1	Belgium Jupiler League	Belgium	2008/2009	1	0	3	KSV Cercke Brugge	RSC Anderlecht	RSC Anderlecht	0	3
492476	1	Belgium Jupiler League	Belgium	2008/2009	1	5	0	KAA Gent	RAEC Mons	KAA Gent	3	0
492477	1	Belgium Jupiler League	Belgium	2008/2009	1	1	3	FCV Dender EH	Standard de Liège	Standard de Liège	0	3

Dataset 2:

The following dataset focus on real player's physical characteristic and FIFA player data

> **Table 3**: Variable descriptions and units for each variable in the dataset

Variable	Description	Units
ID	ID of the Player	NaN
Name	Name of the Player	NaN
Height	Height of the Player	Centimeters (cm)
Weight	Weight of the Player	Pound(lb)
date	Date of the data update	NaN
overall_rating	Overall Rating of the Player	Score from 1 - 100
preferred_foot	Player's preferred foot	NaN

Variable	Description	Units
attacking_work_rate	Attacking Rating	Score from 1 - 100
defensive_work_rate	Defensive Rating	Score from 1 - 100
crossing	Crossing Rating	Score from 1 - 100
finishing	The rating of player finish the whole game	Score from 1 - 100
heading_accuracy	Heading Rating	Score from 1 - 100
short_passing	Short Passing Rating	Score from 1 - 100
dribbling	Dribbling Rating	Score from 1 - 100
free_kick_accuracy	Free Kick Rating	Score from 1 - 100
long_passing	Long Passing Rating	Score from 1 - 100
ball_control	Ball Control Rating	Score from 1 - 100
long_shots	Long Shots Rating	Score from 1 - 100
vision	Vision Rating	Score from 1 - 100

> **Table 4**: Example of relative abundance data.

ID	Name	Height	Weight	date	overall_rating	preferred_foot	attacking_work_rate	defensive_work_rate	crossing	finishing	heading_accu
30723	Alessandro Nesta	187.96	174	2007-08-30 00:00:00	91.0	right	medium	high	42.0	24.0	93.0
30723	Alessandro Nesta	187.96	174	2007-02-22 00:00:00	91.0	right	medium	high	42.0	24.0	93.0
30955	Andres Iniesta	170.18	150	2013-06-07 00:00:00	90.0	right	high	medium	85.0	73.0	54.0
30955	Andres Iniesta	170.18	150	2013-05-24 00:00:00	90.0	right	high	medium	85.0	73.0	54.0
30955	Andres Iniesta	170.18	150	2013-05-17 00:00:00	90.0	right	high	medium	85.0	73.0	54.0

Methods

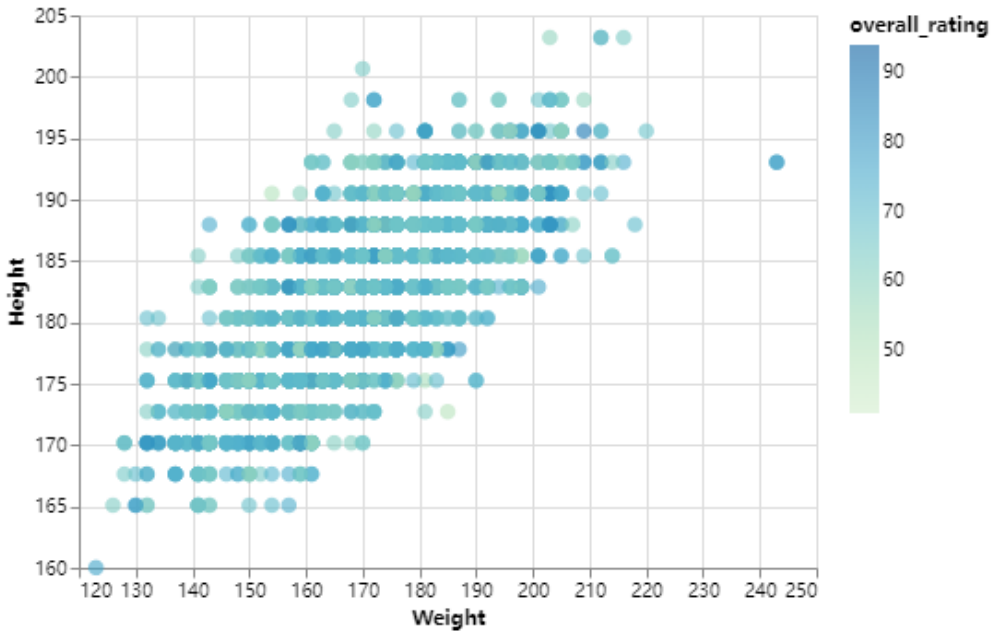
The exploratory analysis aimed at soccer players' physical characteristics and FIFA soccer player data. By applying data manipulation the dataset transferred to focus on the best player in FIFA against a real player's physical characteristics. Subsequently, multiple linear regression was performed, and learning the key features from the data to predict the overall rating by player's height and weight. Also, by applying principal components analysis to simplify datasets and determine the data features to demonstrate the result. In addition, data visualization was made to understand the data better. (Data visualization package Altair was used).

Results

Relationship between height and weight against overall rating of soccer player

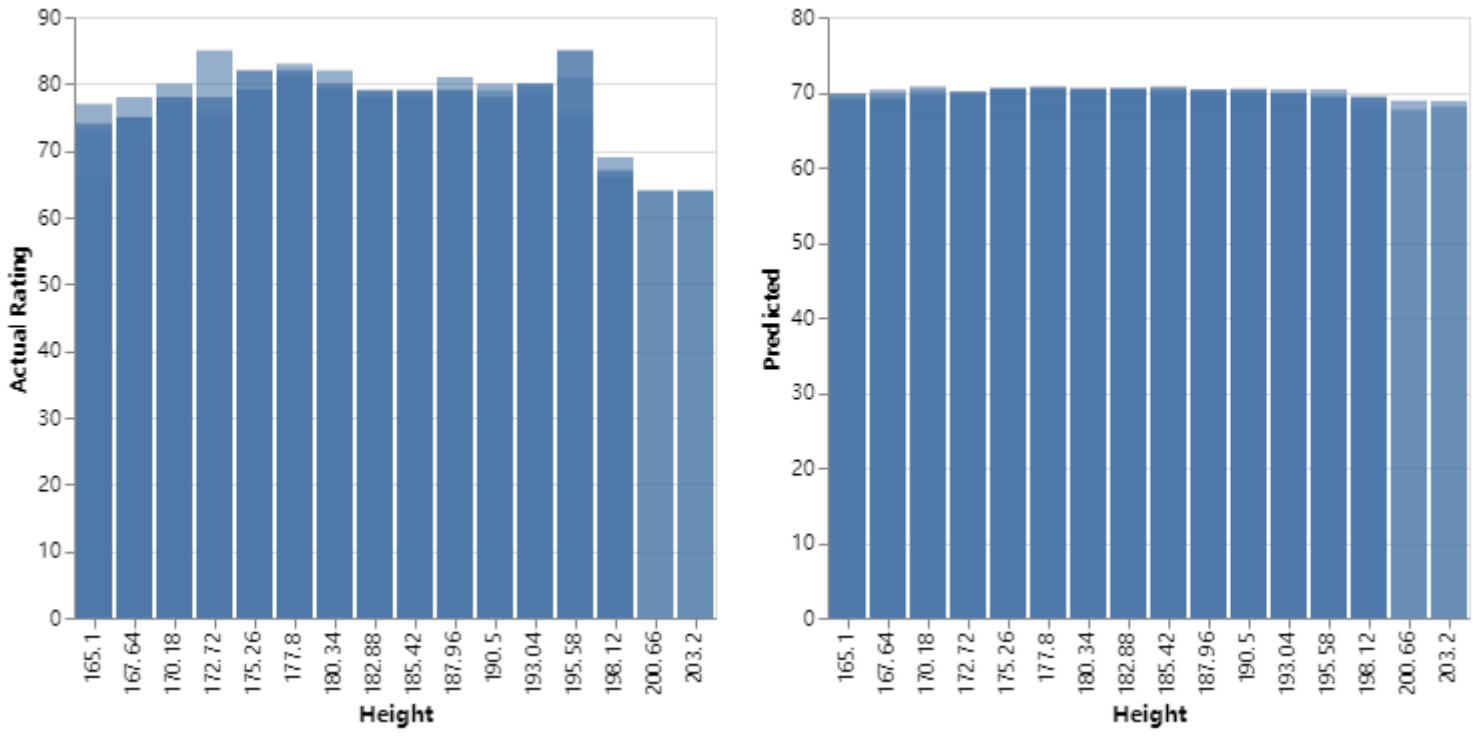
After data manipulation the dataset shows the best player's height and weight. Figure 1 shows the relationship between player's characteristic and overall rating of the player

> **Figure 1**: Figure 1 demonstrates that height and weight are not the key factor that influence the overall rating which means heigh and weight is not providing extremely obvious difference.



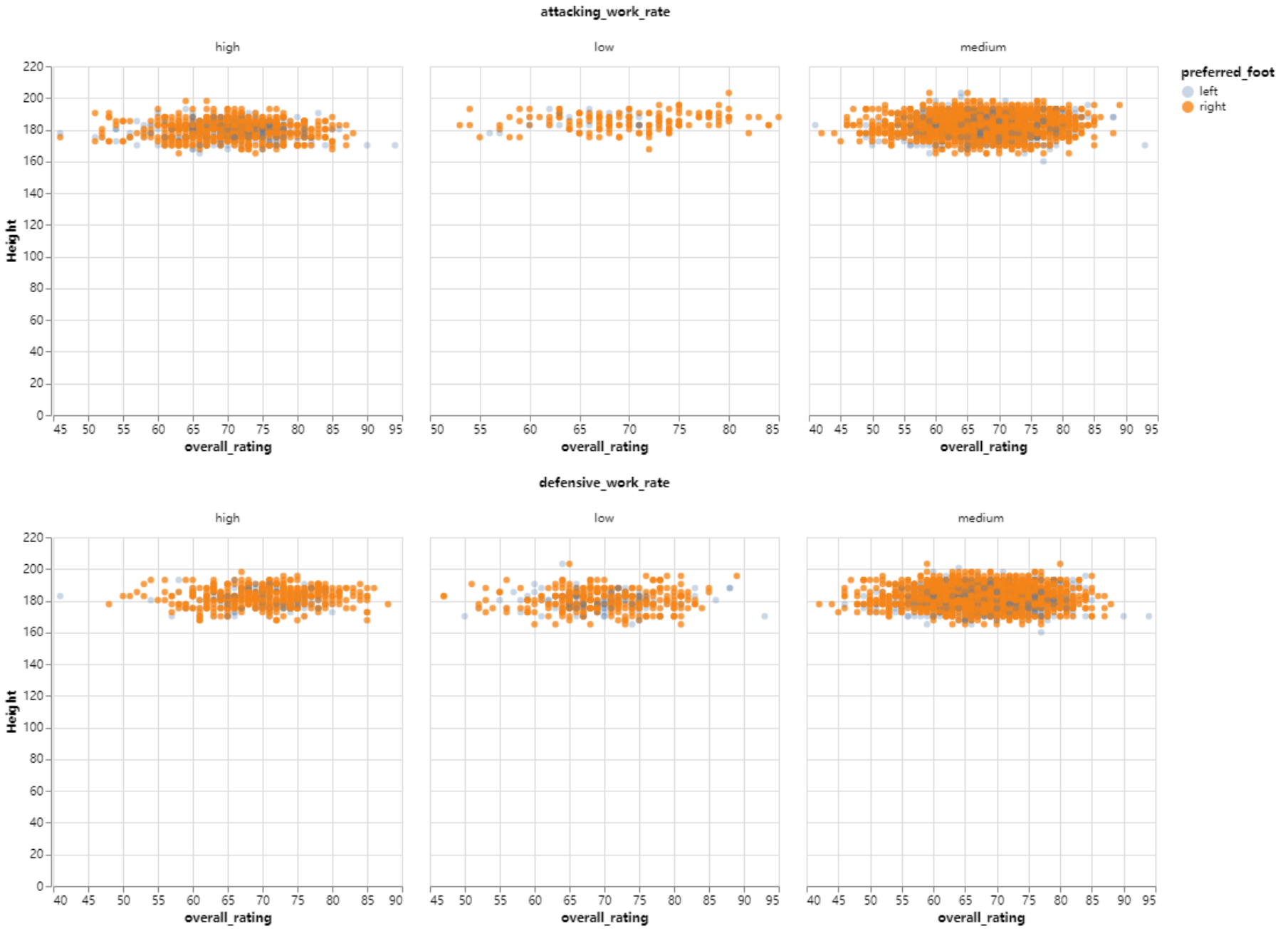
Second, applying multiple linear regression has demonstrated that a height around 175cm will provide a better probability for the player to rate over 90. Even though the data visualization (figure 1) is not showing the difference between height and weight against the overall rating. But after using multiple linear regression to learn the features from data the machine learning model has predicted that 172cm - 180cm height had an obvious difference from other height

> **Figure 2** : Figure 2 demonstrates that whether the real data or predicted data both show that players arround 172cm - 180cm had a better overall rating than others.



Further analysis of how player physical characteristics affect attacking and denfensive work rate during the game.

> **Figure 3**: Figure 3 shows the different level of attacking / defensive work rate of players. By focusing on the high level of both area, the data visualization shows that players with 180cm or above are mostly rated as high in attacking/defensive work rate.



Discussion

This section should conclude your report in 1-2 paragraphs that reiterate the findings and offer any commentary. 'Commentary' could include:

- speculation about the cause of certain findings;
- caveats about interpretation;
- refining of questions or aims;
- further topics you would have liked to explore.

Soccer is a famous sport in the world it has a lot of fans. However, soccer was defined as a sport that requires good physical fitness. The stereotype of soccer would be people thought only strong physical characteristic makes a good soccer player. In fact, physical characteristic doesn't play an important role in how a player performs in a soccer game. According to our dataset, different heights had different advantages which means some of the heights are good at dribbling and some others will be good at free-kick, or passing. Because a soccer game is based on the whole team that means not only the player who got the goal performs well. Good defense and good passing also play an important role in the soccer game. Besides, if we focus on the player's height in certain positions we can figure out a better physical characteristic that has a better win rate. In other words, if we go deeper to compare all the data on the same position we can find out what is the best physical characteristic for each position. In conclusion, we will be able to predict the soccer game result by looking at the player's physical characteristics in each position.