Data Science Project Proposal - Data Science and sports: Fad or Future?

Abhishek Purwaha 46435255
Cameron Warton 44635931
Jack Reynolds 44635206
Zahin Sobhan Enan 46247165

Summary
Quantitative data has always played a role within competitive sports, games are decided by scoring more goals, home runs, and tries than the opposition while players are judged on metrics such as their goal and assist output. However, as technology has advanced, the pool of data to extract meaning from has exponentially increased. A great contemporary example of an organisation successfully navigating this new swath of data can be observed in the book and movie of the same name, Moneyball.

As the relationship between sports and science has become increasingly intertwined since the start of the 21st century, is there such a thing as too much data and analytics? Will data always triumph over the ‘eye test’? Our project explores this conundrum through Football (soccer), with an emphasis on the English Premier League (EPL).

Project Goal
This brings us to the overall question we want to explore, in a world of endless sports statistics, what is the best quantitative statistic for a player’s true contribution to a game?
The most famous sports players in the eyes of the fans have always been biased towards the ones who score the most points for their team. They are seen with a much higher contribution weight compared to the other players in their equally demanding yet less statistically prevalent positions. Traditional statistics like goals scored, tackles, passes, saves are all position-related statistics but aren’t equal reflections of a universal player's contribution. Our overarching goal is to investigate the effectiveness of existing traditional measures versus new arising methods. These new methods may revolutionise the way sports are analysed.

API / Dataset Description
The API contains over 22 different categories of information, categories of particular interest include: venues, fixtures, injuries, predictions, coaches, players, transfers, matchday squad, and betting odds. It will also have data about subscriptions and requests. The sample architecture of the dataset will consist of season, countries, and leagues. It will also have other sub-datasets such as the fixtures, teams, top scorers, standings, labels, and trophies. Dataset about coaches will also be present. The file format is imported as JSON.

Aim / Questions
Easy questions

- What stats are common among past premier league/champions league winners

- Do some countries naturally produce particular types of players?

Medium questions

- **How did Leicester City win the league in 2016? How was it possible for a team that was given no chance to beat the top-flight clubs?**
- Are formations and tactics more influential than player quality?

Hard questions

- The rise of the Expected Threat statistic (xT)
- How good are scouts at finding young football talents? If we take a look at the yearly golden boy awards, how many of these recipients turn out to be world-class players?

Analysis Techniques

We initially we plan to use the following analysis techniques in some form:
Regression – To create a predictive model and compare expected vs actual results
Visualisation - Use of various plots and visual models to showcase our analysis. This will help breakdown our analysis into digestible information that the general public can follow
K means clustering – to help classify our data within logical groups

Milestones

Milestone 1: Week 8

Confirm the modes of communication between team members

- Project Communication to be done via Discord and Facebook Messenger

Confirm the chosen APIs and general direction of data analysis

- Chosen API is from auto-data: www.api-football.com/

Milestone 2: Week 10

Complete data analysis:

- Linear regression
- Recursive Feature Elimination (RFE)
- K-Means clustering
- Appropriate visualisation

Milestone 3: Week 12

Complete delivery of the project

- The project is to be delivered through Jyupter notebooks, which falls under the Anaconda Navigator family
- If there is an opportunity we might deploy our findings to a web app such as Heroku
- Create powerpoints slides and scripts for video presentation