GROUP 87 | Adam Jordan, Saad Abderrazzaq, Gabriel Lam, Kyle Schluns

Harvard CS 109A

Pavlos Protopapas, Kevin Rader, Chris Tanner

13th October 2020

Final Project | Milestone 2 | Scope of Work | EA SPORTS – FIFA

Preliminary Project Statement:

The project aims to understand the performance of players within a video game as a basis of finding a model to explore performance and results predictions in sports. EA's long standing FIFA franchise is an example of how real-world statistics and events have virtual implications. The following questions will form the structure of our investigation into this statement:

- *How do each of the different attributes affect the Overall Score of a player?*
- *Are some more weighted than others? Are players playing their most optimized position, or are their skills better suited for a different position?*
- *How well does a club's staff improve a player's skills?*
- *What teams have the highest level of improvement based on our predictions?*

Through the exploration of these questions via the supplied datasets, we hope to gain understanding towards the interpretation of empirical data; the investigation here may allow us to gain insight to create more efficient and sensitive models that would generate accurate implementations of data into digital products.

Plans for Preliminary EDA:

The workflow for the EDA was based upon the aspects of work outlined on the brief. This would allow us to frame our EDA in a manner compliant to the project criteria. The different components of the project extracted from the Project Guideline are here below:

- Task A: Use data from FIFA 19. Predict the Overall (OVR) skill Statistic for players in the FIFA 20 Edition. Train on all players from FIFA 19. Graphically represent the Overall data for the players in the Test Set.
- Task B: Train on data from FIFA 19. Predict the player_position variable using other skill statistics for players in the FIFA 20 Edition. Train on all players from FIFA 19. Graphically represent the predicted player_position for the players in Test Set
- Task C: Study player data from Division 1 European League* players from the last 5 Years. Analyze changes in player stats and value. Rank the clubs according to the best increase in statistics of a player. Graphically represent the scores for the test set.
- Task D: Going through data for player skill changes, predict the new skill stats for the test set players in the 2020-21 season

We decided to utilize the workflow of the project as a means in which we could begin planning for the EDA required. We were able to find datasets with a creative commons license in Kaggle (https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset) that had scraped from the website prescribed by the project brief (https://sofifa.com).

What is the condition of the data we got? What data do we need?

Firstly, the capturing and cleaning of these datasets are required, as well as the processing of categorical data that is necessary to engage and further exploration; categories such as player_position and preferred_foot have at least one value contained within. In addition, work needs to be done regarding the classification of players and their clubs into their subsequent leagues.

What relationships are we looking for?

Based on the project brief, we can begin to split the EDA into semi-independent components that can be efficiently tackled into more succinct questions:

Part A: Predicting the overall rating of players

- *How do each of the different attributes affect the Overall Score of a player? Are some more weighted than others?*
- *What are the attributes of those with the highest ratings?*
- *How have those skill attributes changed in significance over the last few games?*

This will allow us to begin engaging in specific data points that affect the players and their performance. It will separate data points that are potentially irrelevant to the overall rating of the players, as well as data points that are potentially collinear and therefore requires mindfulness regarding further analysis.

Part B - Player position

- *Are players playing their most optimized position, or are there skills better suited for a different position?*

Presumably, the position will be a predictor of the overall rating regression model from part A. So to test if players are playing their most optimized position, we can run the model prediction multiple times keeping all predictors constant, except change the position each time. If the model predicts a higher overall rating for a position that is different than their current position, then we could conclude they are not playing their most optimized position.

Part C - Which club has the best staff?

- *Does an increase in player spending create an increase in team rating?*
- *Given a player's age and starting skill level, what is the expected improvement?*
- *How does a player's 'potential' affect their future rating?*

To figure out the best increase in statistics, we must first define which statistics are considered valuable to this assessment - for example, to a club, the salary statistic is much more significant to that of a player's personal and individual ranking. We aim to maintain some level of constraint, where increase in statistics can be found purely in the data set without requiring external verification (such as win-loss ratio), but rather the efficiency and 'bestness' of a club is based on the overall change in average team rating with respect to a change in the club spending.

Part D: Predicting subsequent seasons

- *What teams have the highest level of improvement based on our predictions?*
- *How have they differed from predictions based on previous years? Would this team continue its trend of improvement for later years?*
- *Which players have the highest level of improvement based on our predictions? Are they in the teams that have the highest improvement?*
- *Does a player's salary follow their rating? Is it pre-emptive? Is it based on history? Is it unrelated?*

This relies on the models being first completed before predictions can continue, but the framing of these questions allow us to be more open-minded towards the types of predictions that may end up occurring; e.g. teams that improve most may not have the most improved players. These questions also set up future potential explorations that exceed the current scope of this project, and will likely be left as speculative.

Basic Task List:

1. Setup repo/shared environment [Saad] - *Completed*
2. Setup working area/meeting [Adam] - *Completed*
3. Cleanup player data, and processing of categorical data [Adam + Kyle(?)] - by Nov 15 (tent.)
4. Part A [Adam] - by Nov 15 (tent.)
5. Part B [Saad] - by Nov 16 (tent.)
6. Capture & clean-up of team/league data [Gabe] - by Nov 15 (tent.)
7. Part C [Gabe] - by Nov 16 (tent.)
8. Part D [Kyle] - by Nov 17 (tent.)
9. Using results of those parts to answer project questions [Kyle] - by Nov 17 (tent.)
10. Final writeup and submission [All] - by Nov 18

Additional Notes (Per Canvas):

What is your group #?
Group 87

Have you met/communicated with your fellow teammates?
Yes - our primary method of communication is a Slack Channel and Github Repository.

Have you met/communicated with your assigned TF? If not, please provide a reason.
Yes - our main correspondence was with regards to the collection of the data from an existing parsed set rather than creating our own webscraping.