Team Boeing 109

FINAL PROJECT | FIFA 15-20

Group Number - 40

Preamble - EA Sports, the Fifa universe, the dataset and its bias

As there are clubs that are featured. Fifa is not only a soccer simulation with the aspiration to mimic reality to the closest, but also a game that needs to sell. Thus, EA Sports, the company that produces the game, has special contracts with big clubs such as Manchester United, Real Madrid, Chelsea, Bayern Munich, or AC Milan. EA sports seeks partnerships with these clubs, because they have a big fanbase that they can promote the game to. As a consequence, FIFA has less of an incentive to downgrade a player from these clubs and even incentives to overrate players or tweak some stats of a player to make him more playable in game. Thus, in this data set, might be a bias toward bigger clubs as well as to bigger leagues / more prominent divisions. In fact, EA Sports has an official cooperation with the Premier League.

Consequently, all our predictions will inherit this bias. One subtask of all our work throughout the project will be to identify signifiers of this bias and to raise awareness. That said, we believe that our models that are trained with this data set and our predictions can still be of value and relevance within the EA Sports FIFA universe. In addition, we hope to write code and develop methods that can be used on other, real-life soccer datasets to make predictions that didn’t inherit the EA Sports FIFA-bias and can thus be applied more universally.

Part A - Predict Overall Rating

Basic Implementation:

Leverage the entire dataset of all players and train a regression model to determine the overall rating of any player based on their other statistics.

Reach Goal

Come up with more advanced feature selection & develop sophisticated algorithms beyond the scope of e.g. Lasso and Ridge regression to identify the most important features.

Bonus

We expect that the game rates top tier players in a different way than very low level players. Divide players into different tiers/categories and train different models for different tiers.

Part B - Classify player position

Basic Implementation:

Leverage the dataset of all players. Train a logistic regression model to predict the position of players using their other statistics.

Reach Goal

Train a neural network to classify players to positions.

Bonus

In addition to classifying player positions, we plan to also recommend secondary positions for each player. This could be important not only in-game (in case you need a player to fill in a position due to injury, etc.) but even in real life, if you want to provide recommendations for coaches or scouts.

Part C - Which Club has the Best Staff

Reflections:

First step is to identify what features we're looking at. Are we only interested in OVR, overall attribute changes (OAC), or Market values?

From history, we know some clubs had and have a very good staff, from the history of their players and how they develop. We can do some EDA on the players of these clubs to figure out which statistics are increasing, and choose those statistics as our response variable. This chosen response variable would be our numeric meaning of the "Best Staff".

Another challenge we face here is strong incoming or departing players, which wouldn't reflect how good the staff is. To solve this, we only take into account players that stayed at the club in consecutive years. Then, find the difference (improvement or decline) in their response variable (such as overall rating) through years.

Response variable example: If our chosen variable is OVR, then our response variable would be the average change of OVR in a club from a year to another, for players who stayed at the club.

Basic Implementation:

Choose a metric, such as OVR, to be our response variable. Calculate the average changes of the OVR from year to year, and then rank the clubs based on the expected change. Finally, plot the estimated scores on the test set.

Reach Goal:

Try to come up with a model that takes into account the following ideas:

- Based on the dataset, figure out at what age OVRs start dropping. Take that into account when measuring the changes. (e.x.: This player was expected to drop 10 points, but only dropped 5 points…)
- Make use of a new predictor column that compares the yearly change of overall rating and potential overall rating (e.g. ovr - potOvr)
- Since the fifa dataset is biased in terms the numbers players per club (some less prominent clubs even in top divisions have only a handful of players), we might want to play around with 'standardizing' each club before making statements about the best staff (e.g. use only 10 players under the age of 27)

Part D - How will things be in 2021

Reflections

We will make the assumption that the player will still play in the same position in 2021.

Basic Implementation:

Leverage the data set of all players. Train on the changes of a player's skills over the years to predict their changes for the next year.

Reach Goal:

In addition to the predictions of attributes, ovr, and market value, we would like to predict position changes as well as potential transfers. Specifically, we are interested in predicting when a player might leave the club, e.g. he became too big for a club, etc. We would like to do that for both, the dataset including players with expiring contracts as well as for a dataset that excludes players with contracts.)

In addition, we'd like to predict when a player with an expiring contract will sign a new contract with the club.

Ideas for EDA:

For part A: Predict Overall Rating

- Histogram of overall rating.
- Scatterplot for all the players using means/rating as axis (maybe we need to take only a subset of the data set since 17000+ players might be too much to handle for a scatter plot)
- Boxplots or violin plots (such as defenders, midfielders, etc.)
- Seaborn pairplot/ Matplot Scatter Matrix to explore the relationship between attributes.

Part B: Predict Player Positions

- Scatter plot with different colors for each position. Use three dimensions: one for forwards' attributes, one for midfielders, one for defenders. Check for clusters.

Part C: Which Club has the best staff?

- 'clustered' barcharts for all clubs of one division showing changes in OVR, overall attributes, and market value. Based on that graph, enriched and contrasted by historical knowledge, we will be able to select the most suited metric for the main task, which is to evaluate the best staff, i.e. we'll be able to identify whether we will use OVR, overall attribute or market value change to evaluate the best staff.
- In addition, we'd like to get a better understanding of additional key predictors, such as the average age of a team. We are planning to use a scatterplot to encode this information in one graph. Age could be on the x axis, OVR on the y axis, and the radius for the circle could reflect the market value.

Part D: How will things be in 2021

- Choose some features to visualize.
- Explore the dataset for players to identify and eliminate the non-numerical attributes such as Name and URL variables.

EDA Plan

1. Vinit will submit the assignment for Group 40 on Monday at 6 PM
2. Vinit will work on the scatter plot identified above for Part C and present to the team by Oct 31.
3. Vinit will work on the OVR Rating Box Plots and Violin Plots for Part A and share with the team by Nov 7.
4. Robert will do general EDA of the player data to allow for a better understanding of the data overall by October 25th.
5. Robert will do EDA for part B, i.e. create various scatter plots to identify attribute patterns for special player positions.
6. Robert will do final data cleaning by Nov 1st.
7. Masaoud and Ali will work on EDA for Part D.
8. Masaoud and Ali will write the detailed plan for milestone #3 by Nov 17/18.