Chris Galdi
CPSC 392

Final Project #3

Beginning of the Analysis:

A little intro about FIFA 21 and the beautiful game that it encompasses. I want to talk about some of the major leagues around the world, as well as some of the relevant national teams and players that have made a large impact in the football (soccer for the noobs) world. I also want to talk about FIFA itself, explaining some of the aspects of the game across the many years as well as the game dynamics, ratings, and other in-game things. I want to use this section to explain some of the purpose of why I am looking into the data and what it represents overall.

First Question:

What is the spread of nationalities of the players on FIFA? Which nationalities have the highest amount of highly rated players? Are there any nations that have very low membership in FIFA? Which position has the highest number of players? How does each league compare to each other? What are the most populous leagues in FIFA? How does it compare to the known Top 5 Leagues (Premier League, Ligue 1, La Liga, Bundesliga, and Serie A)? Which league is the most valuable and filled with the higher overall players?

a. For this question, I want to begin by visualizing the data, looking at the spread of values for each league, country, club, nation, rating, value, and position through the usage of scatterplots, pie charts, box and whiskers, and bar graphs to spread out the data to explore the different players. I am going to be looking at the count of the league, country, position, nation, value and club to see where some of the players play, looking at the spread of players, which nation has the most players in certain positions, giving a gauge on the location of where these players play and which leagues are the most populated with players and from which teams.

b. I am looking to first get a look at the spread of data overall and presenting that view in the form of a question and an in-depth look into the spread of the players will give a good overall visualization of them and the respective players of the stats that we will be looking into in the next questions. Will the most populous leagues and the leagues with the most money and value have an overall advantage over other smaller, homegrown leagues? The correlation between the leagues and the players is what makes talking about football interesting and gives the Champions League and Europa League some backing in terms of why teams qualify.

c. As I am looking to first get a look at the spread of data overall, I feel presenting that view in the form of a question and an in-depth look into would be adequate visualizations before doing other analysis. Being able to see how the data relates and the different spread of the players correlate to the leagues overall and how well (or bad) a team might perform on FIFA and in real life.

d. I will definitely be using scatterplots and bar graphs as my visualizations for this question, as well as box and whiskers and pie charts. The question is based around visualization and trying to look at the spread of players, not clustering or predicting these as they are ever-changing random variables, so using these visualizations will show the spread of data and how leagues compare to each other, etc.

Second Question:

Can we predict the overall rating of a player? How can it relate to the actual rating on FIFA?

a. For this question, I plan on using a Linear Regression model with the predictor variables of the stats (pace, shooting, physical, strength, dribbling, defending), age, weight, wage, release clause, and value to predict the player. I will be utilizing KFold with 10 folds, as well as z-scoring my variables and calculating the mean squared error and the r squared score for model validation. I will also be calculating the coefficients and odds coefficients for the predictors to see which predictors had the most effect on the data if shifted.
b. I think these choices are good because they focus on the values that FIFA considers when trying to create a rating and determine the value of the player, as well as being a lot of aspects of a normalized scouting report. Z-scoring the variables puts it on an accurate standard scale and leaving 10 folds gives it enough data in each fold to have a good sized spacing for the amount of data.
c. These will answer the question by providing viable predictors to predict a rating, while also using a Linear Regression model given that the rating would be considered continuous.
d. I will be using scatterplots and bar graphs to look at the spread of the data as well as box and whisker to look at the spread of the data, as well as scatterplots to visualize the true vs predicted values.

Third Question:

Can we cluster the in-game stats (attacking_*, defending_*, skill_*, movement_*, power_*, mentality_*, defending_*, goalkeeping_*) and game stats (pace, passing, shooting, dribbling, defending, and the goalkeeper ratings) to help group types of stats and what type of player that person may be?

I am going to make a clustering model with both of the stats being used as the predictors. Then, I am going to create two clustering models, one for the in-game stats and one for the game stats to group the players. From there, I will look at both of the models and make observations and then compare those to the initial larger model. I am doing this in an attempt to see if they can be clustered and to see how the in-game stats or game stats relate.

a. I am planning on using Hierarchical (Agglomerative) clustering method to cluster the two separate stat attributes. I am going to z-score the variables, attempting to look for the best

number of clusters but looking to start at 4, as there are 4 main groupings of positions when roughly categorizing a player's position. I am going to use the aforementioned in-game stats and game stats as the variables, as they in theory encompass the rating of a player. I will be using silhouette scores to determine how well the method fits the data with the clusters given, as well as to determine how well the clusters fit the data overall.

b. I chose to do it this way because I wanted to see the relation of the stats and whether or not they can actually be used to show a player's position or if the stats could be just attributes.

c. The HAC method will help to show how the structure of the different stats that are given to the player and how the stats can be clustered to determine how a player will play, as well as being able to compare the two stats to determine which one of the two better dictates how a player will play in game.

d. I will be utilizing scatterplots and dendrograms to visualize the data to show the grouping of the clusters as well as the relationships of the clusters to the different stats.

Fourth Question:

Can the in-game stats (attacking_*, defending_*, skill_*, movement_*, power_*, mentality_*, defending_*, goalkeeping_*) be used to predict overall rating? Can the normal game stats shown on the card (pace, passing, shooting, dribbling, defending, and the goalkeeper ratings) predict the overall rating? How do both compare to each other and the true rating given by FIFA?

I am going to construct two different models, one with strictly the in-game stats, and the other with the normal stats. From there I will compare and evaluate how each did.

a. I am going to be using Linear Regression models to predict the overall rating for both in-game stats and game stats and utilizing the LASSO regression for dimensionality reduction. I am going to use the aforementioned in-game stats to predict the overall rating, as well as the overall rating given by FIFA. I am not going to remove outliers, but I will be z-scoring the variables. I will be using a starting cv of 5 and getting the mean absolute error for the data. I am going to be using two different models, so I will have two sets of predictors to

b. I think that these are accurate measures because I want to find a way to accurately predict the rating and feel that Linear Regression is the best way to do that to determine if the ratings effect the players or if the stats are just numbers for the game and the introduction of the LASSO method will enable dimensionality

reduction and should help give more accurate results, as well as giving me the ability to look at a smaller amount of columns to predict.

c. These will answer the question of if the stats can actually predict the actual value or if the stats don't really have an effect on the rating that FIFA gives players each year.

d. I am going to use scatterplots and histograms to support my answers, using them to show the relationship of the predictors as well as the predicted vs true values and overall just using them to compare the data and visualize the model results. I will also be using boxplots to look at the individual stats.