Data

The Fifa player dataset was downloaded from Kaggle, and contains all the soccer players included in the videogame scraped from the 2019 Fifa video game itself.

Data Wrangling

Dataset from Kaggle contains 18206 rows and 89 columns. I performed some data cleaning and data transforming on the following features:

- Removed 75 rows to keep the important variables for my research (the variables kept are described below)
- Minimized the rows to only the top 1000 players and removed the players with missing information
- Converted the “Positions” variable categories into 4 main groups A, GK, M, and D
- Transformed the “Height” variable into a binomial
- Transformed the “Preferred” variable into a binomial
- Cleaned the column “Value(M)”
- Cleaned the column “Wage(K)”
- Cleaned the column “Weight”
- Created the column “Over27”

Outlier data points were looked at individually and determined if the number was an incorrect entry or a legitimate measurement, and corrected the incorrect entries. At this point, we have 991 rows and 12 columns left in our cleaned data.

VARIABLES

Name: Name of the player (Categorical)
Age: Age of the player (Continuous)
Over25: whether or not a player is younger than 27 (Binary: 0 no, 1 yes)
Nationality: Nationality of the player (Categorical)
Overall: Overall rating of the player from 1-100 (Continuous)
Potential: Potential rating of the player 1-100 (Continuous)
Club: Soccer club the player plays on (Categorical)
Value(M): The players value in euros and millions (Continuous)
Wage(K): The players wage in euros and thousands (per week) (Continuous)
Preferred: Dominant foot (Binary: 0 left, 1 right)
Position: Soccer position (Categorical: A: attacker, GK: goalkeeper, M: midfielder, D: defender)
Height: whether or not a player is above 6 foot (Binary: 0 no, 1 yes)
Weight: players weight (Continuous)

Questions

1. Which variables (Height, Weight, Overall, Potential, Age, Position, or Preferred) have the strongest impact on a player's value?
2. Is the model used to answer question 1 better at predicting the accuracy for players younger or older than 27?
3. Can we predict what position a player plays based on their weight and whether or not they're taller than 6 foot?
4. Is the relationship between wage and overall stronger than the relationship between value and potential?
5. Do players that are left footed and above 6 feet have a higher potential rating than players that are less than 27 years old and shorter than 6 feet?
6. What is the relationship between a players overall rating and value and is the relationship different between those two variables for players that are are and aren't over 21?
7. Do players on club teams in Europe make a higher wage on average than players on club teams in South America?

Analysis Plan

Question 1:

Which variables (Height, Weight, Overall, Potential, Age, Position, Preferred, or Club) have the strongest impact on a player's Value?

a) For this question I will be using both linear regression and dimensionality reduction. Because this question is aiming to explain which variables impact a player's value the most (which is a continuous variable) and we are assuming that there is a linear relationship between our variables, linear regression is the algorithm most appropriate to answer this question. Our dependent variable will be the player's value and the independent variables are Height, Weight, Overall, Potential, Age, Position, Preferred, and Club. When using this algorithm we must Z Score the continuous variables in order to standardize the values. In addition to linear regression, to answer this question we will be using dimensionality reduction and more specifically Lasso. Lasso will take into account possible noise in our data and tell us which variables that we've chosen don't have an affect on the outcome of a player's value. When using Lasso we must choose a lambda value which I will choose based on looking at the standard deviation of error of the variables and choosing the standard deviation value that produces the least amount of error.

b) After running our model we will be able to use both MSE and R Squared to predict how well the model does in predicting a players value as mentioned. If the R2 is very high (close to 1) then we know that the model is highly accurate however the lower the R2 score the worse the model is and the less linear relationship there is. If we have a high R2 score then we can assume this model is very good at predicting a player's value and it would be a good implementation for the company. In terms of evaluating the MSE, there is no clear way to determine what a good MSE is because it depends on the variables and each dataset therefore when we get our MSE we will evaluate it based on our model specifically.

c) To support the conclusion of the implemented algorithm I will be using two ggplot graphs to depict the predicted players value against both the actual players value and the error. The first graph (predicted v actual) will provide a visual for us to see the relationship between the predicted and actual values of the players value. The second graph (predicted v error) will be used to present a visual of how correct/incorrect our model is. We can use this graph to test for homoscedasticity and normality of errors.

Question 2:

What is the relationship between a players overall rating and value and is the relationship different between those two variables when taking into account Age?

a) For this question I will be using a clustering analysis algorithm to show patterns and groups between the two variables overall and potential rating and later adding the variable age to see how it affects our groupings. The clustering method that I will be using is DBSCAN. The reasoning behind this selection is that I don't want to assume any shapes of clusters because the relationship between the variables doesn't indicate any clear shapes as well the fact that I am only clustering on three variables and DBSCAN does a good job with fewer variables. Additionally, DBSCAN accounts for noise which will give us a clearer visual and representation of actual groups within our data set with the possibility of having outliers. When using DBSCAN there are two hyperparameters to be chosen which are epsilon and minimum number of neighbors. In order to pick the value for these metrics I will use a number of inferences surrounding our data set. For example, for the minimum number of neighbors I will take into account the number of players in our data set and pick a number of neighbors proportional to it, because we only have 1000 players we know that we want a relatively small number of neighbors. In addition, we don't see an overwhelming amount of noise therefore this further explains why a smaller number would work more adequately for the data set. For epsilon (the distance we will look for these neighboring points) I will create a K distance graph and choose a value at the elbow of this graph.

b) Clustering is the best form of answering our question because we aren't looking for the specific outcome of any one variable rather for patterns or groups that can be created from the data. There are a number of clustering algorithms that could work to solve this question however DBSCAN makes assumptions about data sets that are concurrent with our data set and has unique features which would benefit us in terms of result output. We can use a silhouette score after clustering to see how strong the clusters are and later compare the original clusters with the clusters when taking into account age. This will allow us to see if the age of a player has any affect in the grouping of potential and and overall rating of players.

c) In order to provide a visual for our clusters there are a number of graphs that will be made. For the first clustering of just the overall and potential ratings of a player there will only be one graph representing their clusters by color. Once we add the element of age we will have to reproduce graphs and there will be a total of 3 additional graphs. These graphs include potential v overall rating, potential v age, and overall v age. We have to create three graphs because we want to be able to visualize the clusters and because we cant 3d plot them we have to output every combination of variables together.

Question 3:

What is the probability of players that are left footed and above 6 feet being older than 27 compared to players that are right footed and shorter than 6 feet?

a) In order to answer this question I will be using probability metrics. Because we have two distinct groups of people left footed and above 6 feet and right footed and below six feet we will be calculating probability for both of these groups to see what the probability is that these players are older than 27. When looking for probability I will use groupby and later create a data frame to allow us to make visuals for our conclusions.

b) This analysis choice works best because we will be getting the output of probability which can be compared between the two groups to explicitly answer our question to see if these probabilities are different or similar.

c) One of the graphs I will be using to visualize the results is a bar graph of the two groups and their probability score to see the difference and similarity more clearly. Other graphs will be used as well before the outcome and is going to be the relationship between the variables in the data set. For example the relationship between age and dominant foot as well as age and height.