INTERMEDIATE REPORT

FIFA PLAYER RATING AND WAGE ANALYSIS

SUMMARY:

The project proposes to find the various features/skills that affect a player's overall ratings in FIFA. As a first step, the data was cleaned and normalized to produce a consistent dataset that can be used for dimensionality reduction and clustering. Since the ratings depend on different set of skills based on the position of the player, the dataset was partitioned into four based on the position of the player. Dimensionality reduction was performed using PCA to identify the important features in the partitioned dataset. Then, we performed clustering on the whole dataset to identify groups of players based on their ratings. The resulting clusters were visualized in a 2-dimensionsal space using pairs of important features obtained from PCA to obtain vital insights about the data.

DATA PRE-PROCESSING:

The dataset consists of 18.2k rows and 89 columns where each row represents a player and each column represents their attributes/skills. The dataset contains both textual and numeric fields. The following challenges were observed with the dataset:

- Columns like Wage, Height and Weight had to be normalized.
- Missing values in the dataset.
- Many columns irrelevant to our analysis.

Primarily, the missing values in the dataset were filled with the mean of the entire column. The wage, height and weight columns were normalized to similar units as the skill-based columns (on a scale of 100) so that we maintain consistency while calculating distances/similarities. All the normalized columns including skill set, positional score, weight etc. were converted to int64 format for performing PCA and clustering. Some highly incomplete rows were completely removed from the dataset to reduce the noise. Moreover, some of the columns that didn't contribute to our objective like Club logo, Club Name, Joining date etc. were also removed from the dataset to keep it precise and reduce computational costs. All of the cleaning/normalization on the original dataset was performed on the data frame obtained from Pandas.

We have a total of 34 skills available for each player like Crossing, Finishing, HeadingAccuracy, GKDiving, GKHandling, GKKicking etc. There are total of 27 positions available in the dataset such as 'CAM', 'CB', 'CDM', 'CF', 'CM', 'GK', 'ST' etc. Each player has a score for each position and each skill (out of 100). We mined this data to find out the top 10 relevant skill set needed for each position played. We analyzed the data based on the relationship between skills and positions. Using insights from the data, we found out the top 10 skills that contributed the most to each of these 27 positions.

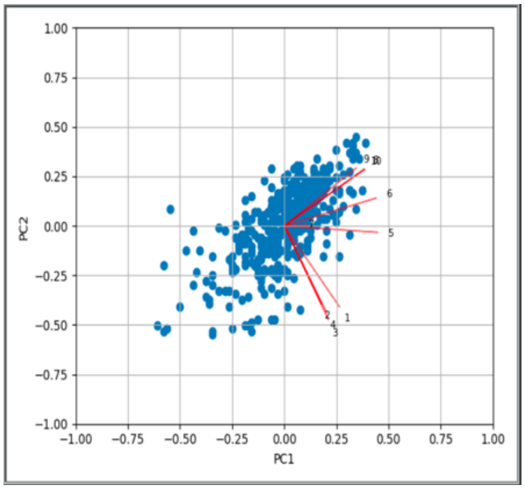
We then mapped each position to 4 broad categories like ‘Attack’, ‘Defense’, ‘Midfield’, ‘Goalkeeping’. We aggregated the skillset obtained previously and found out the top skills for each of these 4 positions that the players take in the field as below:

- Goalkeeping: GKReflexes, GKDiving, GKPositioning, GKHandling, GKKicking
- Midfield: Balance, ShortPassing, Agility, Stamina, Acceleration, SprintSpeed, BallControl, Dribbling, Jumping, Aggression, Strength
- Defense: SprintSpeed, Acceleration, Stamina, Agility, Balance, Jumping, Strength, Aggression, StandingTackle, HeadingAccuracy
- Attack: Agility, Balance, Acceleration, SprintSpeed, Dribbling, BallControl, Jumping, Finishing, ShotPower, Positioning

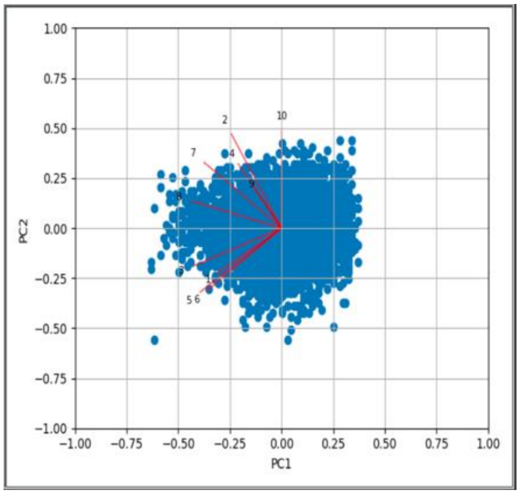
It was noticed that these top 10 skills contributed more to the overall ratings than the rest.

DIMENSIONALITY REDUCTION (PCA):

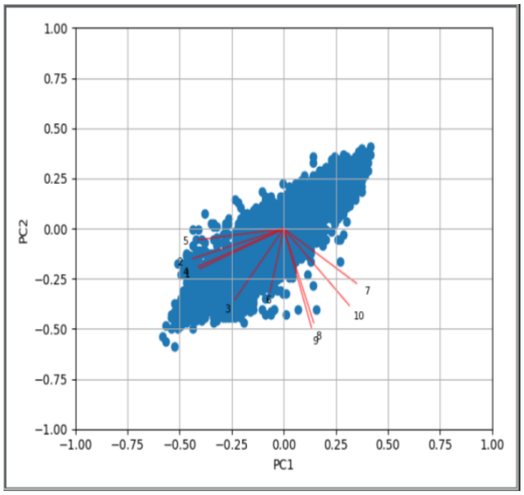
Prior to performing PCA, the data was normalized with StandardScaler such that its distribution has a mean value 0 and a standard distribution 1. This ensures that PCA gives a fair comparison between the explained variance in the dataset. PCA was run on the four partitioned datasets since we need to find the features with high variance for each of the four categories. Each category had its own dataset with top 10 features identified from the data preprocessing step. We ran PCA on the 4 datasets to find the two most important features for each position. PCA was run with two components and the resulting principal components were plotted on a biplot which depicts the magnitude of the variance among the features as below.



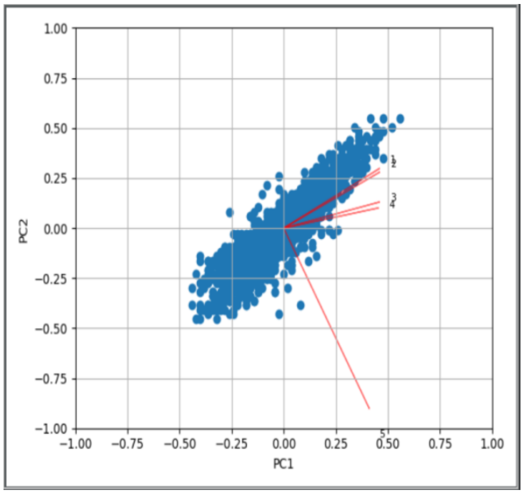
Attack



Midfield



Defense



Goalkeeping

Each red line in the plot represents the eigen vector for each feature and the length of the line represents the magnitude of the values in the eigenvector. More the magnitude of a feature, the more important that feature is to explain the variance of the dataset. Based on the magnitude analysis from the biplot, we choose the following two features as important for each of the four categories:

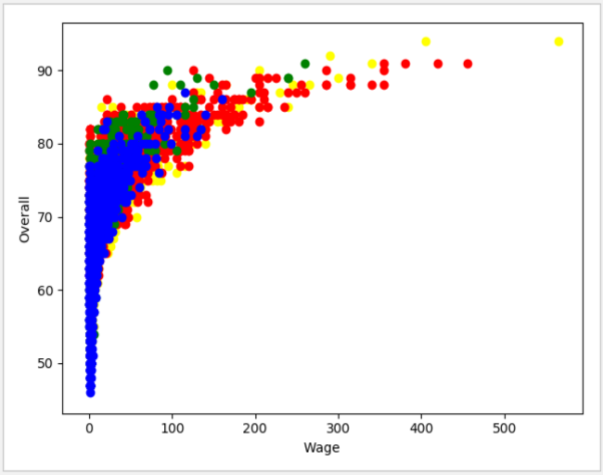
- Attack: Dribbling, BallControl
- Midfield: Jumping, Aggression
- Defense: - Strength, HeadingAccuracy
- Goalkeeping: - GKReflexes, GKDiving

CLUSTERING:

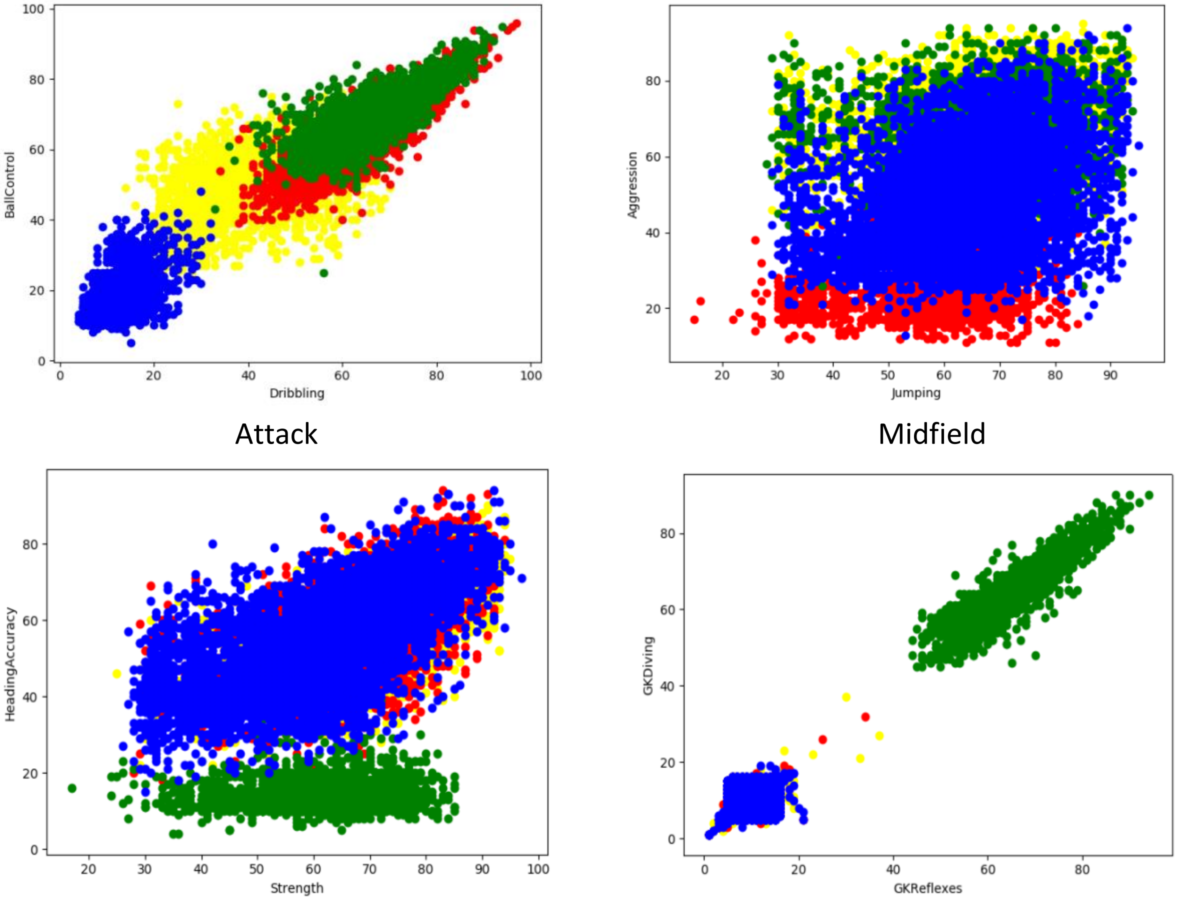
We cluster the dataset to obtain various insights which would help us understand the correlation between the features better. We have chosen the number of clusters to be 4 (k=4) based on the broad positions available in the football field. The four positions are namely, attack, defense, midfield and goal keeping. We only use the columns representing the skillsets to cluster the players. K-means ++ clustering is performed and plotted against the overall ratings and wage as shown in the figure below.

We could observe that though two players have similar overall ratings, the player in attack position is paid more compared to the defender or goal keeper. We could infer this intuitively from real world statistics that attackers score goals for their team and hence have more brand value in the market. Hence, they are paid more compared to others. For all the subsequent graphs, the colors representing the categories are:

Yellow: Attack, Red: Midfield, Blue: Defense and Green: Goal keeping.



Further, we also plotted the clustered data against the top 2 important features obtained for each category from PCA. The plots show that the players with higher values for the 2 features fall into the corresponding cluster representing their category i.e. Attack, Midfield, Defense or Goal keeping. This can be observed across all the plots below. Further, analysis of the plots indicated that the higher the values of these features in each category, higher the overall ratings of the players were.



Attack

Midfield

Defense

Goalkeeping