Analysis of FIFA'17 and FIFA'18 Data

Omkar N. Kulkarni, Vishnu Sharma, Gabrielle Giacoppe, and Zhiyong Li

AMAT 502 - Modern Computing for Mathematicians,
University at Albany, State University of New York, Albany, NY, USA
Email: {onkulkarni,vsharma3,ggiacoppe,zli34}@albany.edu

I. Introduction

In the modern data driven world, it is becoming natural to fuse analytics with a variety of different domains. One of those emerging fields is Sports Analytics in which statistics, computing, and the right analyst can derive pertinent insights from every facet of the game. This allows one access to information from the game previously denied simply because no one was looking. The movie 'Moneyball' is an excellent example of it, where data analysis in baseball was introduced back in 2011. Due to the abundance of data that is being generated nowadays by the advent of technology, analyzing it has also become important. In that spirit, we decided to perform some comparative analysis on the FIFA 17 [1] and FIFA 18 [2] data sets from Kaggle. In this analysis, we examine these two data sets from the popular FIFA video game by EA. The aim is to implement multiple machine learning methodologies like principal component analysis due to the presence of numerous features, regression analysis on a continuous target, and clustering after exhaustive feature engineering. It is important to pursue this endeavor for several reasons. First, if insights can be derived by fusing two seemingly separate fields and presented as a deliverable, it leads to action, which is an accomplishment. Furthermore, in scientific processes, when one attempts something new it broadens the scope of what can be done which is exciting for future experimentation that can build on previous findings. Data holds meaning no matter the sector. Therefore, manipulating it and extracting insights drives innovation and that includes sports, and, in that spirit, we explore these data sets.

A. Machine Learning Insight

There are a handful of applied machine learning techniques applicable to comparative analysis; a complete guide to how they covered can be found with their depths in the further sections.

Chosen methodology, and when to utilize them, include:

- `Principle Component Analysis`: Suitable for dimensional reduction by discovering variables that are most correlated and describes the variation of the data sets.
- `Clustering`: Partitions elements of attributes which are based on player features and whether they were offensive, defensive, or neutral in order to narrow down a non-overlapping Offensive and Defensive classification to then calculate the overall score.
- `Regression`: Displays a correlation regarding a dependency relation, resulting in the distinction of the response variable, whose output is determined upon independent player attribute inputs.

B. Assumptions

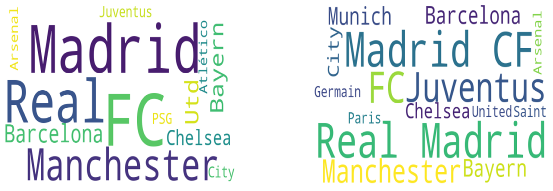
In this analysis, as it is a video game, we assume that gamers of the same level play the game. Hence, the results do not consider gamer's performance but only concentrate on the player statistics. We would also like to point out that as this analysis is based on video game data, the results may not always reflect the same in the real world. Although, this work can provide the basis for anyone interested in taking it up for real-world soccer or any other sports data.

II. Data Cleaning and Preparation

In this analysis we made use of the two publicly available data set from the popular website Kaggle in which data science enthusiasts from around the globe congregate to share, contribute, and collaborate on several data science projects. Both the FIFA data sets were uploaded to this website by its users. The data sets contain player statistics for the soccer video game created by EA Sports for the years 2017 and 2018. They both contain over seventeen thousand values and over fifty features. To be exact, the FIFA'17 data set contains 17341 values with 53 features, whereas the FIFA'18 data set contains 17793 values with 75 features. Both the data sets have a total of 25 common attributes or features between them. As with any data science task, it is important to discern noise with relevance. Therefore, data cleanup process takes precedence over everything where we needed to make decisions on the useful data and noise.

Data cleaning is the process of detecting and correcting values from a data set and refers to identifying irrelevant, duplicate, incomplete, inaccurate or incorrect values and deciding on whether to remove them or modify them. As expected from the size of our data sets, there were a lot of values and features or attributes that needed cleanup before we proceeded with the analysis. Some attributes in the FIFA'17 like National Position,National Kit, Club Position, Club Kit, Club Joining, Contract Expiry were irrelevant to our analysis, and needed to be dropped from the data set. Similarly the FIFA'18 attributes like Unnamed: 0, Photo, Flag, Potential, Club Logo, Value,Preferred Positions, Wage were also ignored. Along with dropping some features, we also had to drop some rows that had inconsistent values and inappropriate data types. For instance, one of the rows had the value of 40+8 in the Agility feature where the intended datatype was an integer value. Furthermore, in the FIFA'18 data set, all the attributes except for the Age, were of the Object datatype. So, one of the other cleaning task was to convert all the attributes to the appropriate data types.

The FIFA data sets contain player based statistics; the aim of analysis was team based, and hence we had to group the players based on their teams. As it can be guessed form any sports statistical data set, almost every player has two teams that he plays for: one is the national team and other is the club team for other leagues than the national ones. In this analysis, we focused only on the leagues performance of teams, meaning that we grouped them based on the Club attribute and not the Nationality attribute. Once finished with the cleanup, we decided to work on exploratory data analysis on the data set to get more insight about the data sets.



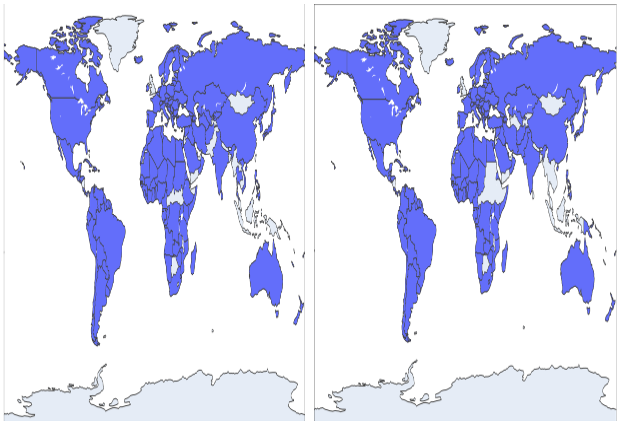
(a) Word Cloud FIFA'17 (b) Word Cloud FIFA'18

Fig. 1: Word Cloud

Figure 1 represents a world cloud for the top 10 teams based on their appearance in the data set. This effectively gives us the top 10 teams that have most number of players in them. We used this in order to get a sense of how the teams and the number of players are distributed between the teams. Figure 1a explains the distribution in FIFA'17, whereas figure 1b explains the distribution in FIFA' 18.

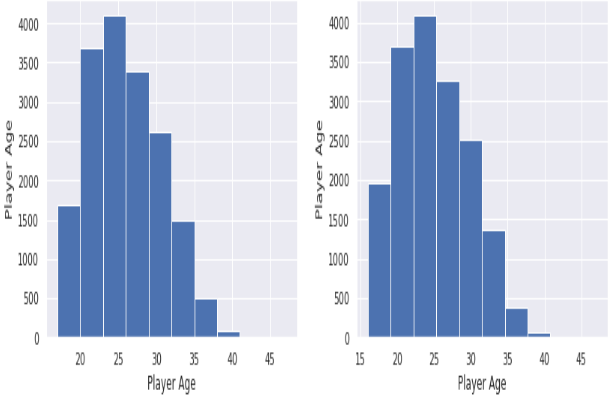
In figure 2, the geographical distribution of the players is depicted. As we can see, all the players in FIFA' 17 and FIFA' 18 are distributed all over the world. Countries like Greenland, Indonesia, Philippines and PNG do not have any players who play in the dataset. We can also observe that there is a slight change in distribution of players from countries like Sri-Lanka, Sudan and Ethiopia as we move from FIFA 17 to FIFA 18, as can be seen from figures 2a and 2b respectively. This can also be because of the retirements and other factors of the real world soccer data that dries the developers of the game to make the changes accordingly.

Figure 3 represents the distribution of players based on their age for the FIFA'17 and FIFA'18 data set. As can be observed in the figure 3a, maximum number of players raged from the age of 24 to 26 and the same trend was observed along the FIFA'18 data set as well, as shown in figure 3b.



(a) Geo FIFA'17. (b) Geo. FIFA'18.

Fig. 2: Geographical distributions



(a) Age FIFA 17 (b) Age FIFA 18

Fig. 3: Distributions based on Age

III. Principal Component Analysis

Principal Component Analysis, also known as PCA, is a dimensionality reduction technique used to condense numerous features to a few components while retaining the largest amount of explanation of variance as possible. In order to move ahead with the PCA pipeline, the data set must be stored in a mxn matrix.

$$D = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{1,n} \end{bmatrix}$$

After completing the required data cleaning, PCA can be implemented. This process begins with mean centering of data because that ensures that every value is on the same scale so there are not any outlier influences.

$$\vec{\mu} = \frac{1}{m} \sum \vec{x}_i = (\mu_1, \mu_2, \ldots, \mu_n)$$

Re-centering of data is done by subtracting the mean from the data matrix. The next step is to build the covariance matrix and retrieve its eigenvalues. They are then ordered from highest to lowest, which are known as the principal components that explain how much variance is explained per component. Lastly, after identifying the number of components that explain the most variance, those components are interpreted and projected onto a two-dimensional or three-dimensional space.

In our analysis, after implementing PCA, it was found that two components explain approximately 80 percent of variance. This can be visualized with a Scree Plot. The Scree Plot is the number of principal components plotted against the eigenvalues. The plot begins in the top left corner and declines steeply before plateauing. The instant it starts to plateau, it is an indicator of the most explanation of variance [3]. In figure 5, we can see that the plot begins to plateau at 2 components. Another important note is that the although the x-axis begins count at 0, it is still being counted by 1. Here in both instances, we can state that two components are the most important.

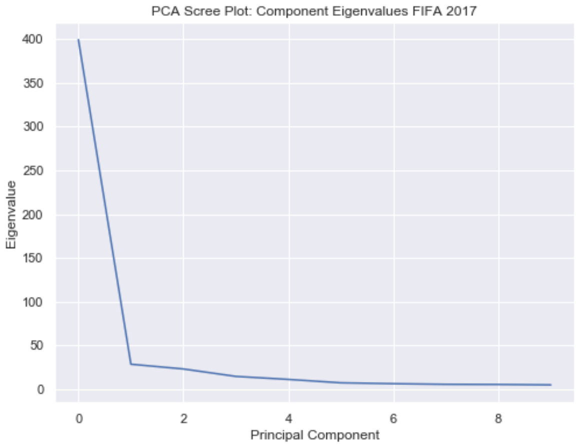


Fig. 4: Scree Plot for 2017.

By utilizing PCA, we were able to take the initial 35 features at the onset of our analysis and condense it down to two components that explain approximately 80 percent of variance. A tidy data set is important to carry out the PCA pipeline properly. Therefore, looking beyond this analysis, we can state that further cleaning of the data set as well as broadening our focus to three components may lead to a better result from PCA. With three components, the projection would be in a three dimensional space which may establish a better understanding of the relationship between components compared to a two dimensional projection.

IV. Projections of PC1 and PC2

The projection of the approximately 600 groups are displayed for both FIFA 2017 and FIFA 2018. There are no discernible clusters in the two-dimensional scatter plot. FIFA 2018 is approximately Gaussian which can be seen with in figure 7.

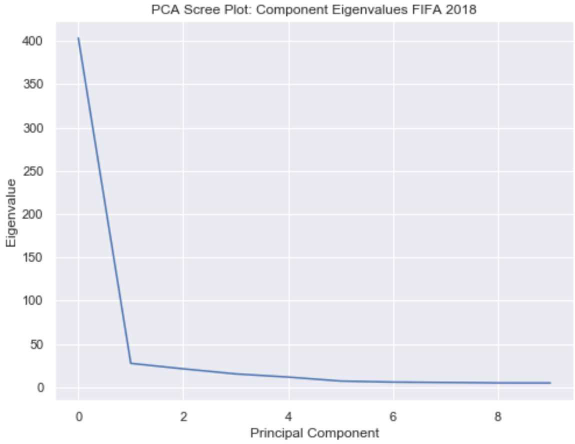


Fig. 5: Scree Plot for 2018.

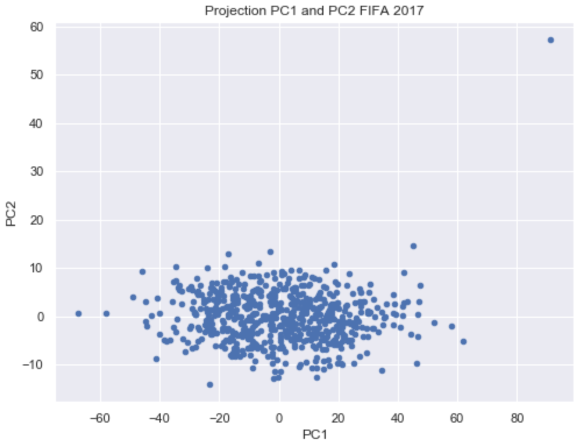


Fig. 6: Projection of PC1 and PC2 for 2017.

V. Clustering

Clustering can be defined as the task of grouping a set of objects or values in such a way that every item in a group (or cluster) are similar to each other in some way [4]. A K-Means clustering algorithm is a constrained optimization problem where it is constrained by the number of clusters and also constrained to try and minimize the sum with cluster variation formula. In our analysis we have made use of the K-Means clustering algorithm for grading the teams based on the number of wins that they achieved in those years in pursuit of a two-fold goal. One, to even out the odds of a poorly performing team playing again the top teams. The other, to promote creation of new leagues of scheme where the teams play only those teams that are close to their performance,

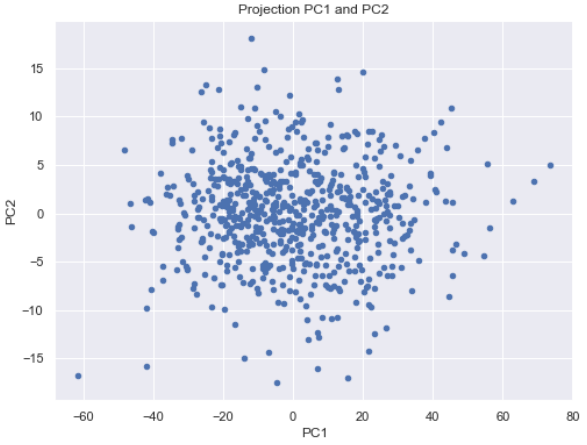


Fig. 7: Projection of PC1 and PC2 for 2018.

minimizing the gap of rankings between them and hence making it more interesting.

A. *Offensive and Defensive Scores*

In this analysis, it was decided to go with 5 clusters based on the number of wins per team. But, as explained in section II, the data was already grouped by their clubs or team names. Each player had about 25 to 30 attributes that were needed to be incorporated in making a decision of synthesizing new feature. Two new features: 'Offensive Score' and 'Defensive Score' were synthesized by making use of these multiple attributes. In the FIFA'18 data set, features used for calculation were: Acceleration, Aggression, Agility, Balance, Ball control, Composure, Crossing, Curve, Dribbling, Finishing, GK diving, GK handling, GK kicking, GK positioning, GK reflexes, Heading accuracy, Interceptions, Jumping, Long passing, Long shots, Marking, Positioning, Reactions, Short passing, Shot power, Sliding tackle, Sprint speed, Stamina, Standing tackle, Strength, Vision, Volleys. Whereas in FIFA' 17, features like Skill Moves, Ball Control, Dribbling, Marking, Sliding Tackle, Standing Tackle, Aggression, Reactions, Attacking Position, Interceptions, Vision, Composure, Crossing, Short Pass, Long Pass, Acceleration, Speed, Stamina, Strength, Balance, Agility, Jumping, Heading, Shot Power, Finishing, Long Shots, Curve, Free-Kick Accuracy, Penalties, Volleys, GK Positioning, GK Diving, GK Kicking, GK Handling, GK Reflexes. These attributes were manually classified as being either offensive attributes, Defensive attributes or neutral attributes depending on their interpretation in the real world.

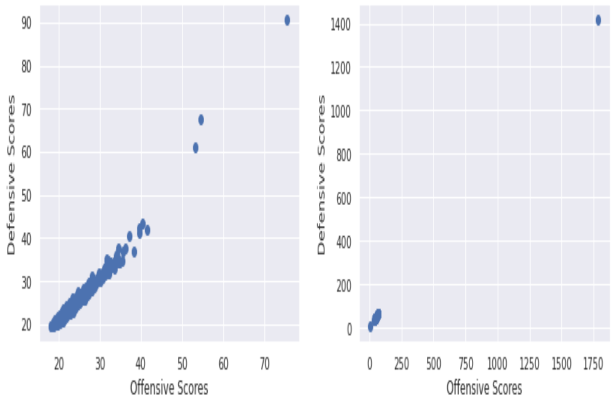
The offensive score was calculated by using equation 1

$$Offensive\_Score = \frac{Mean(\alpha)}{n} \tag{1}$$

Where α is Mean of Sum(Offensive Attributes, Neutral Attributes) and n is the Total number of players in the team. The defensive score was calculated by using equation 2

$$Defensive\_Score = \frac{Mean(\beta)}{n} \tag{2}$$

Where β is Mean of Sum(Defensive Attributes, Neutral Attributes) and n is the Total number of players in the team.



(a) Offensive vs defensive scores FIFA'17.

(b) Offensive vs defensive scores FIFA'18.

Fig. 8: Offensive Score vs Defensive scores.

As can be seen in the figure 8 where we plot the offensive scores vs the defensive scores to see a correlation between them and they seem to be closely related. Apart from a few outliers, all the values seemed to be lying between 0 to 100 in FIFA'17, whereas between 0 to 45 in FIFA'18.

B. *Results Simulation and K-Means Clustering*

With each team now associated with their own offensive and defensive score, we proceeded to simulate the results for each team playing against every other team. The aim was to simulate the results for all the teams playing against every other team in order to get the total number of wins for each team when played with all the other teams in the data set. This resulted in a combination of factorial(n) results. The results were decided using the simple comparison of the offensive and defensive scores. Any team with the maximum of both- the offensive and defensive score was assumed to win the game between the two. In a case where both the competing teams had either of the offensive or the defensive score higher than the other, the result was assumed to be a draw. At the end of this simulation, we counted the number of wins for each of the teams and that was the basis for our k-means clustering algorithm. A manual ground truth was created with the following cluster labeling rubric:

- Cluster A: If a team has won more than 550 games
- Cluster B: If team has more than 450 and less than 550 wins
- Cluster C: If team has more than 300 and less than 450 wins
- Cluster D: If team has more than 100 and less than 300 wins
- Cluster E: If team has less than 100 wins

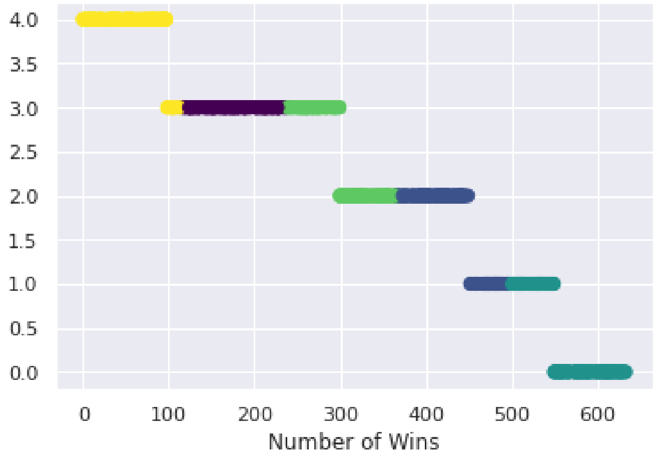


Fig. 9: K-Means FIFA'17.

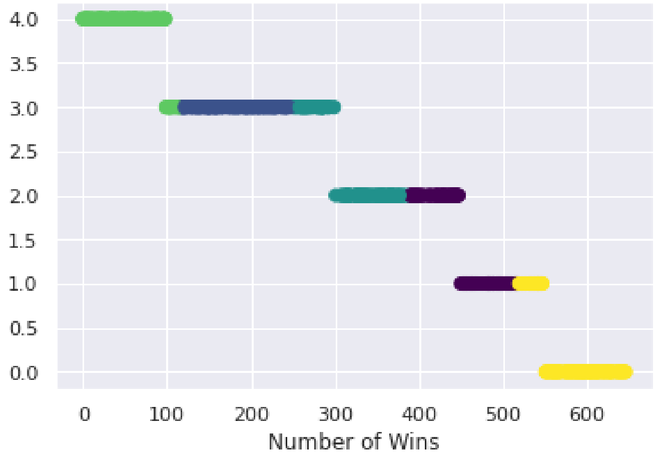


Fig. 10: K-Means FIFA'18.

Figures 9 and 10 depict the K-Means clustering algorithm performed on the FIFA'17 and the FIFA'18 data set respectively. As we can see, the K-means clustering algorithm was quite successful in clustering the teams based on the number of wins. The purity of the method was calculated using the following formula:

$$purity = \sum_i \frac{N_i}{N} p_i \tag{3}$$

where, N_{ij} is the number of objects in the cluster i that belongs to class j and p_i is calculated as:

$$p_i = \max_j p_{ij} \tag{4}$$

Using equation 3, we found that the k-means clustering method had a purity of 75.5% for the FIFA'17 data set, whereas for the FIFA'18 one, the purity came out to be 77.12%. The accuracy was around 18% and 23% for FIFA'17 and FIFA'18 respectively when the accuracy function provided by the sklearn library was used. We concluded that this low accuracy was attributed by the manual choice of ground truth, and the selection of umber of wins to classify each team into. This can be improved in the future efforts.

VI. Regression

Continuous player rankings located within the columns of the data set respective to each club contributed to the placement for predictive modeling with Simple Linear and Multivariate Regression in order to predict the response variable, Shot-Power. Nearest to discovering which attributes the response variable was closely related to, in search of an initial dependency relation, a test of correlation was administered to ensure there was a proper "fitting" within the relationship. First, Vision was declared independent, and Shot Power outputs, or known as the y-values, were influenced based on the x-values. It is concisely positive in slope, demonstrating that as Vision moves to the right on the horizontal axis ranking towards 80, that Shot Power will then travel greater along the vertical axis. This is verifying that enacting linear regression on FIFA'17 and '18 will give desired results for that same predictor. The following depictions of these correlations are shown via scatter plots.

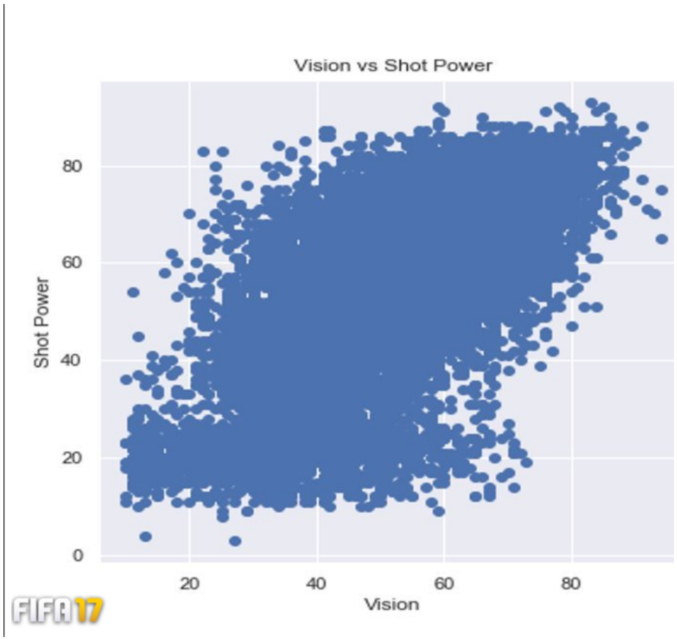


Fig. 11: A visual correlation of FIFA's 2017 Vision with Shot Power.

Table I is read considering the collection of all possible x and y coordinates without inducing the practicality of Train Test Split, which will be discussed later in the section.

It encompasses the coefficient of determination referenced in notation as R-squared, and encloses in the best fit line of regression contents (i.e. slope, intercept); the first row corresponds to FIFA'17 leaving the second related to '18. The line of best fit corresponding to the table is accounted for in Fig. 13 and 14

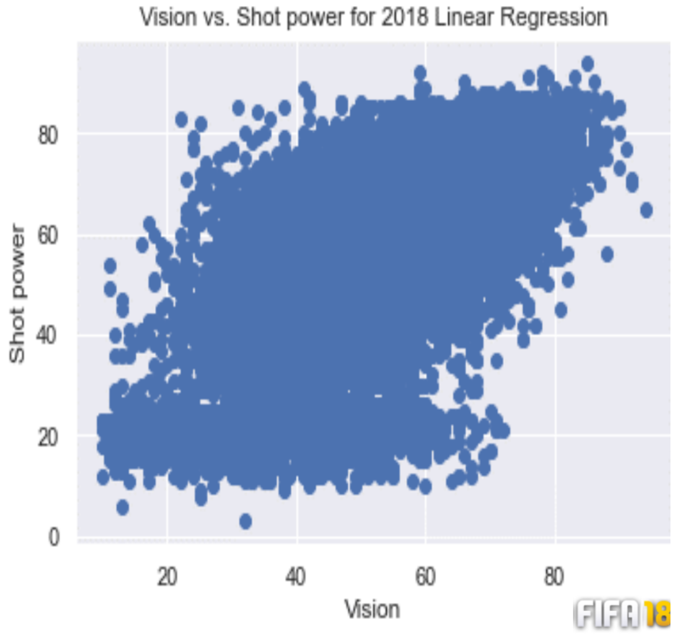


Fig. 12: A visual correlation of FIFA's 2018 Vision with Shot Power.

TABLE I: Simple Linear Regression for '17 and '18

Slope	Intercept	R^2 Score
0.81946	12.38899	0.46143
0.80394	13.02096	0.44253

Train Test Split enables assistive programming technologies to recognise 75 percent of the data usage for training, while the remaining 25 percent was to be tested on. This overcame the pitfall referred to as over-fitting, and guaranteed that identical inputs would not be replicated. The R-squared numeric for 2017 was 0.47342 and 2018 was 0.4519. To compare these values with previous versions, from 0.46143 to 0.47342 and 0.44253 to 0.4519, both having a slight increment, indicating a likely direction.

VII. Multivariate Regression

Of numerous features including Vision, the original hypothesized explanatory variable that would enable significant correlation with the response variable were Long Shots, Composure, and Strength. Strength was prioritized to be relevant. However, after running T-test diagnostics, the evaluated p-value equaled 0.64, much larger than 0.05, deeming this feature ineffective. Since it lacks distinction to use in our model, it was dropped. To confirm these findings were appropriated, see below. Consequently, there is no pattern, and it cannot be concluded that there is any relation between Strength and Shot Power due to the intensity of the randomness in the locations of the points.

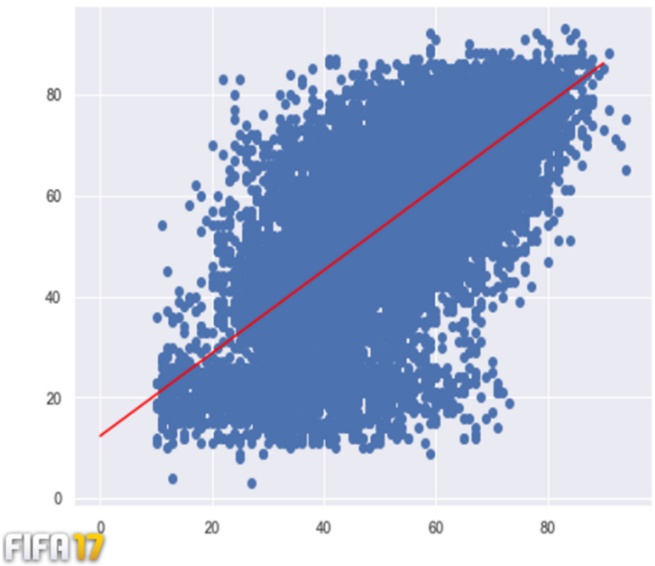


Fig. 13: A visual representation of the best-fit line correlation of FIFA'17 Vision with Shot Power.

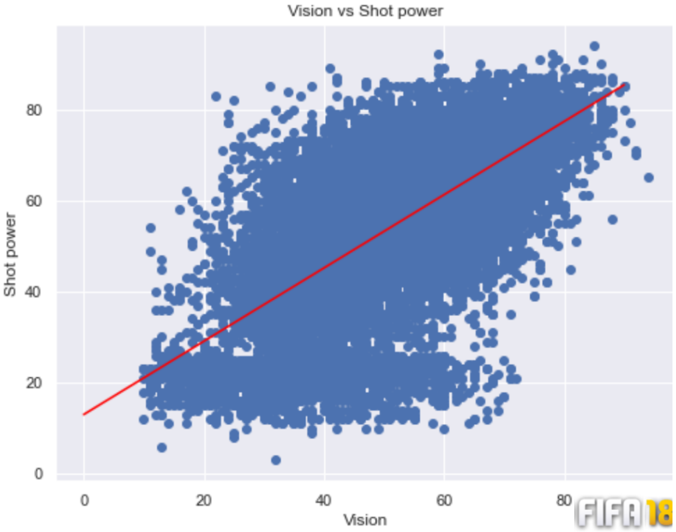


Fig. 14: A visual representation of the best-fit line correlation of FIFA'18 Vision with Shot Power.

Now running multiple linear regression,notice the difference, how the R-squared value constitutes a degree of success at 0.79. While an R-squared to be 1.0 showcases a perfect fit by the predictors, thus far, in comparison to one-dimensional, simple linear regression, multiple dimensional features have given a most plausible coefficient of determination.

Again, to attain a coefficient of determination that surpasses the development of three features, albeit fitting the two models simultaneously, in either linear regression environment, seeking additional player characteristics is beneficial to exploration of Club dynamics consisting of the player's skill level.

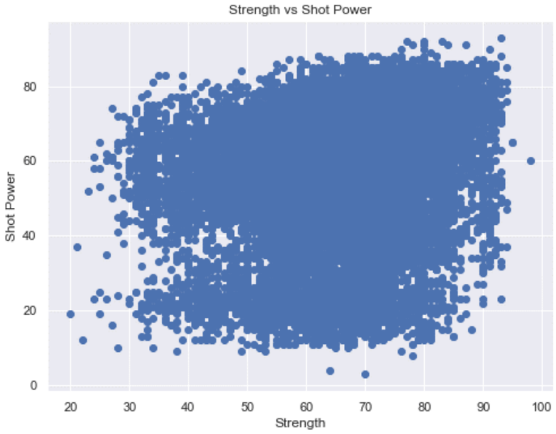


Fig. 15: A representation of no clear correlation of Strength with Shot Power.

VIII. Conclusion

A remark for closing is the acquirement of domain knowledge of EA's FIFA video game series installment. This topic has allowed its authors to become well-versed in cleaning data file pairs, heightening the dimensional pursuit of a multi-dimensional discussion; all to showcase the enhancement of addressing machine learning methods: PCA, K-Means Clustering, and Regression. Their exploration within the maximum allotted time frame of this strategic scope delivered a great medium to the comparative analysis. An initiative to the future is to tweak efficiencies in the data cleaning, assessing those data conditions that were dropped, and ensuring that all possible significant file components were exhausted. The goal being to continue making progress in machine learning, and reevaluating any scientific human errors that may have been impacting the forecasting in this report.

Principal Component Analysis, like several other machine learning tools, rely heavily on data cleaning and preparation. As such, if the analyst is lacking domain knowledge about a particular programming language or package, it may hinder the ability to prepare the data set properly which may impact the overall model. Therefore, for future reference, once more technical knowledge is acquired, PCA can be expanded to more dimensions and perhaps better relationships can be determined.

At first, we postulated Free-Kick Accuracy as a reasonable target variable since it was largely prevalent to all of the features, leading it to be considered as the outcome variable, rather than Shot Power. However, it was changed because of unfeasible entries nested in FIFA 2018's Free-Kick Accuracy column of the newly cleaned data-frame. Otherwise, the collection of delineated features throughout file pairings remained uniform in magnitude. In no circumstance, did the entries undergo scaling, leveraging a critical acclaim in consistency buried in the game design's verisimilitude.

As we all know, in data science data cleaning is the most intensive as well as most important step in the process With K-Means we can partition the given data into k-clusters based on the means. This clustering can be useful in squaring out the odds of weaker teams playing against teams that are significantly better than them. It can also be a basis for new leagues that take the team's performance into consideration. I.e., teams with less than 100 wins can play against teams in their own cluster. With more time on the analysis, some comparative analysis with varying clusters and models with higher purity can be tested for insightful results

IX. Acknowledgments

Tasks completed to make this report come together were collaborative, as each one of us contributed: Omkar cleaned the 2018 dataset for exploratory analysis, composed Word Cloud for the top teams, and K-means clustering to grade the team skill levels of which he elaborates on in his writing of that portion. Vishnu cleaned 2017 dataset to then conduct PCA on 2017, working on the scree plots with projections of which he discussed in writing that section. Gabrielle aided with writing the code for Regression alongside Vishnu, and the 2018, working with Omkar to get def./off. attributes into the equation format, and wrote the Regression section. Zhiyong crafted a speech for the introduction of this report and the presentation.

References

[1] S. Agarwal, "Complete FIFA 2017 Player dataset (Global)," https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global/data?select=FullData.csv, 2017, [Online].

[2] A. Srivastava, "FIFA 18 Complete Player Dataset," https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset, 2018, [Online].

[3] IBM, "Scree Plot," https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=reduction-scree-plot, [Online].

[4] W. User, "Cluster Analysis," https://en.wikipedia.org/wiki/Cluster$_{a}nalysis, [Online]$.