**Predicting Potential of players in FIFA**

**1. Introduction**

Football has a huge fanbase all over the world. Every year, after the end of season, the transfer window allows different clubs to buy, sell or loan players to/from other clubs. Also, the betting agencies start placing their bets on whether a player will complete his rumored transfer to the club in picture. Before spending millions on a player, the clubs analyze the player’s past performance and his future potential to make sure if he will be worth the hassle or not. EA Sports’ FIFA 19 is the latest version of their football simulation game. FIFA provides ratings of the players based on the performance in the past season, and his potential based on attributes like passing accuracy, dribbling, crossing, finishing, height, weight, etc. Different clubs would certainly want to predict the potential of the player before finalizing a deal to get him to their club.

1.1 **Problem**

The aim of this project is to be able to predict the Potential score of a player based on the data present in the dataset. We also want to inspect what attributes factor into determining a soccer players Potential score. The dataset contains the details of players, their nationality, and other attributes such as dribbling, acceleration, stamina, shot accuracy, etc.

1.2 **Interest**

Football clubs around the world want in-depth analysis before putting in a bid for the player in question. The scouting teams from different clubs’ scout players extensively before recommending a player to the club. The clubs would, therefore, be very interested in predicting the potential of a player before buying.

**2. Data Source**

The players’ data for FIFA 19 can be found on kaggle.com. The complete dataset was downloaded if form of a CSV file. This dataset contains the players’ details with attributes that would be useful in predicting the Potential score of the player.

**2.1 Data Cleaning**

The dataset contains complete details of the players attributes such as age, preferred foot, weak foot, wages, skill moves, crossing, finishing, stamina, header accuracy, shot accuracy, etc. Some of the attributes such as stamina, strength, acceleration have a few null entries. These null entries have been replaced by the mean of the attribute to remove any discrepancy.

**2.3 Feature Selection**

A few attributes such as body type, face, flag, etc. have been removed as they will not be used to predict the potential of the player.

The following features have been chosen to predict the potential based on correlation heatmap:  
1. Age

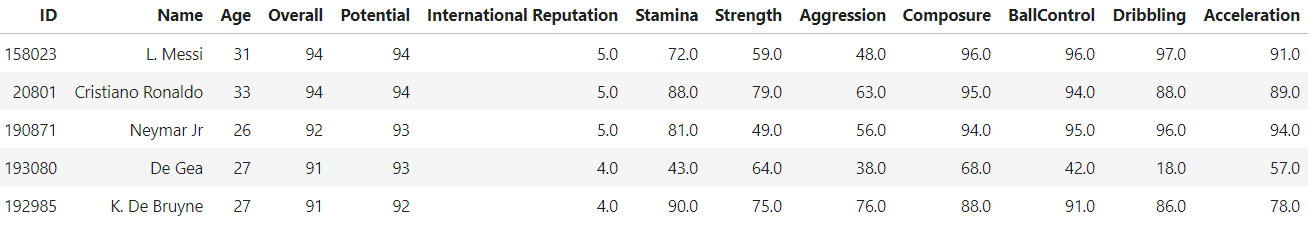
2. International Reputation

3. Stamina  
4. Strength  
5. Aggression

6. Composure

7. Ball Control

8. Dribbling  
9. Acceleration



**Chart, treemap chart

Description automatically generated**

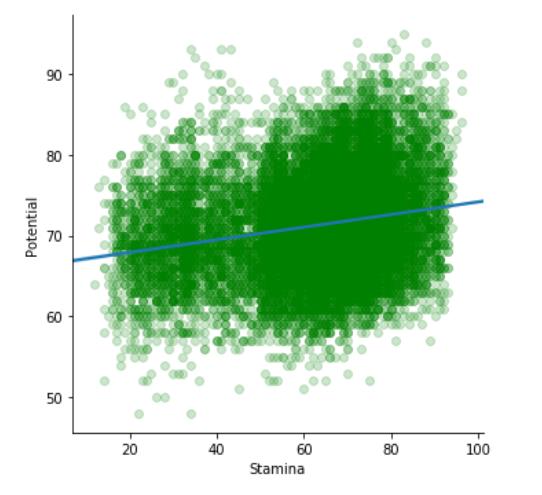
**3. Exploratory Data Analysis**

**3.1 Target Vari1able**

The potential of a player has been chosen as the target variable. The potential of a player represents how a player would perform keeping in view that the player remains injury free for most the duration of the season.

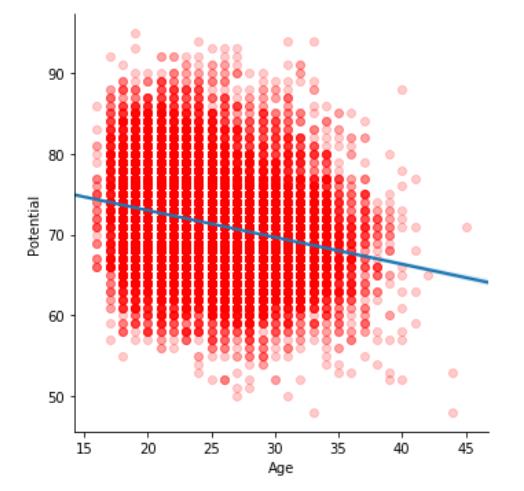
**3.2 Potential & Stamina relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Stamina’ attribute. ‘



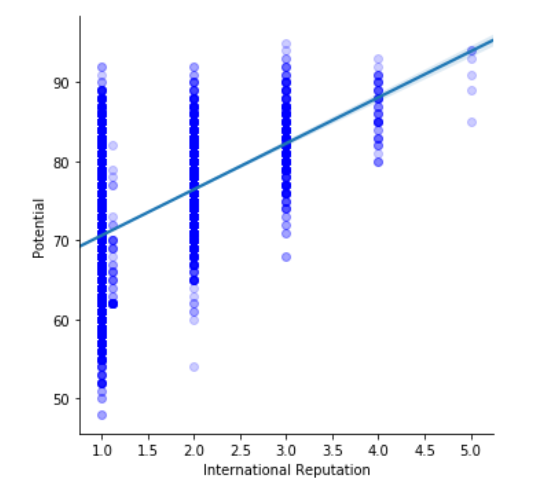
**3.3 Potential & Age relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Age’ attribute. The older a player gets, the less potential they have. This logically makes sense.



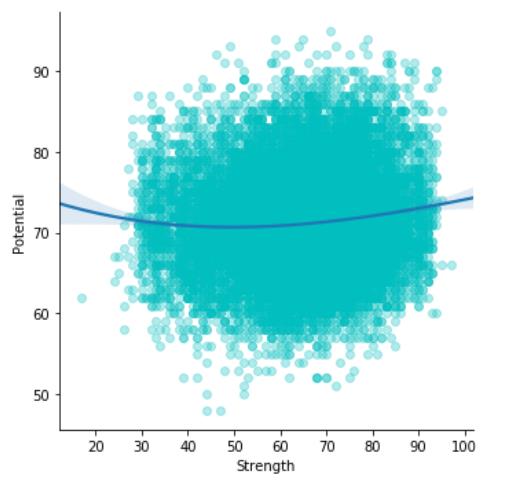
**3.4 Potential & International Reputation relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘International Reputation’ attribute.



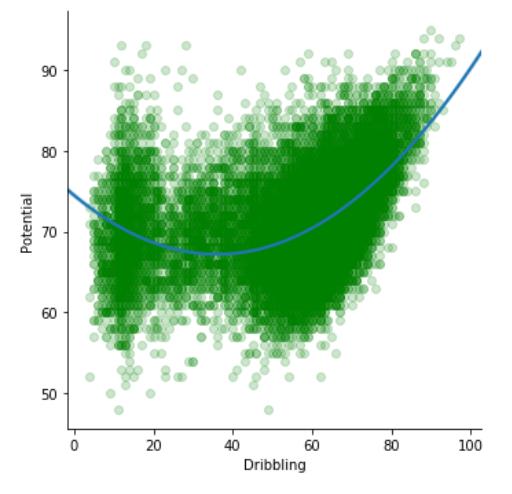
**3.5 Potential & Strength relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Strength’ attribute.



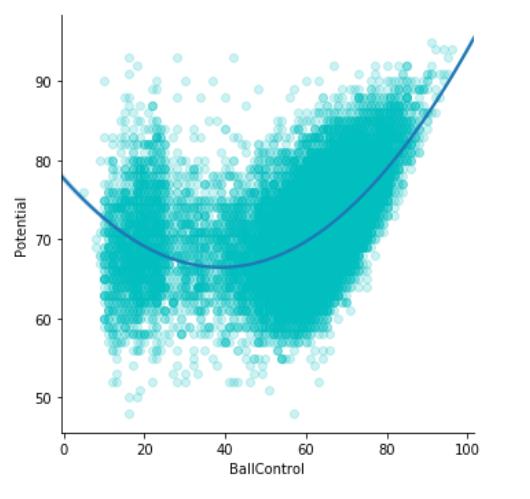
**3.6 Potential & Dribbling relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Dribbling’ attribute.



**3.7 Potential & Ball Control relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Ball Control’ attribute.



**3.8 Potential & Overall Rating relationship**

The following scatter plot shows the relationship between the target variable ‘Potential’ and ‘Overall’ attribute

Chart, scatter chart

Description automatically generated

**4. Predictive Modelling**

I have used Regression models to predict the potential of a player based on other attributes present in the dataset. I start off using a Basic Linear Regression only using one independent variable. Later I implement further regression models such as Multiple Regression, Decision Tree Regression, have been used to predict the Potential and their accuracy has been compared.

**4.1. Basic Linear Regression Model:** For Linear Regression, the attribute ‘BallControl’ has been chosen as the independent variable, and ‘Potential’ as the target variable as defined above. I chose BallControl because it was the highest correlated soccer statistic to potential player score. The following joint plot shows the potential with respect to BallControl.

Chart

Description automatically generated

The following table shows the Intercept, Coefficient and the Mean Squared Error of the Linear Regression Model:

Text

Description automatically generated

Clearly, using just the highest correlated soccer statistic, BallControl, does not give our model great accuracy when it comes to predicting a player’s potential. This makes sense because there are several factors that go into determining Potential. Let’s add more features to our model and see if there are any improvements to our other models

The following attributes have been chosen as the independent variables to find the target variable:

|  |  |
| --- | --- |
| **Independent Features** | **Target Feature** |
| **International Rep.**  **Stamina**  **Strength**  **Aggression**  **Composure**  **Ball Control**  **Dribbling**  **Acceleration**  **Overall** | **Potential** |

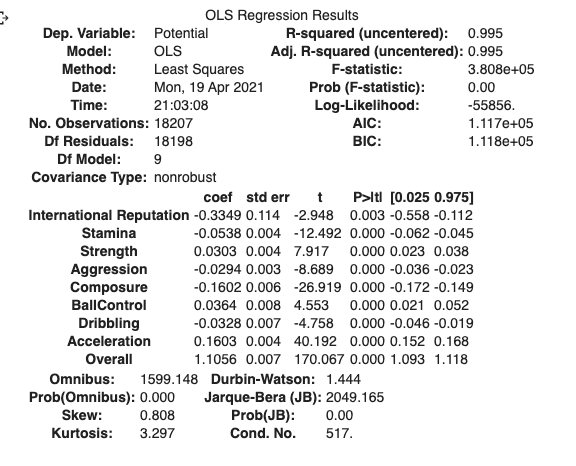
**4.2. Multiple Regression Model:**

**Chart, scatter chart

Description automatically generated**

**The OLS Regression Summary is shown below:**

The adjusted R-Squared means that 98.6% variables are explained in the dependent variable. The t-values show that all the variables are significant and none of them can be ignored.



**4.3. Decision Tree Regression Model**

**A picture containing chart

Description automatically generated**

**4.4. Random Forest Regression Model**

Graphical user interface

Description automatically generated with low confidence

**4.5. K-Nearest Neighbors Regression Model**

**Graphical user interface, application

Description automatically generated**

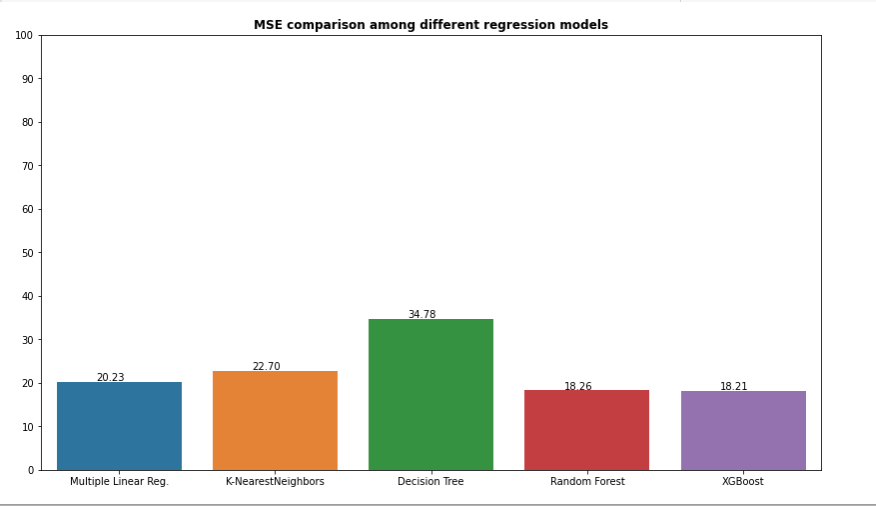
**4.6. XGBoost Regression Model**

Graphical user interface, application

Description automatically generated

**5. Conclusion**

After looking at all our models, I would conclude that using either the Random Forest Regressor model or XGBoost Regressor model would be the best in predicting Player Potential. These models have the highest accuracy and lowest MSE out of all the models we tested. Even though the highest accuracy score we achieved was approximately 53%, accuracy isn’t what is important here.



**6. Discussion**

**6.1 Further Modeling**

This model can be further analyzed using clustering algorithms to create clusters of players with a certain potential. For example, players with a potential greater than 95 can be clustered into a ‘special’ category, while players with potential between 90 and 94 can be categorized as ‘exciting’, and so on.

More complicated soccer statistics are very hard to find without having membership access to them. Being able to access those statistics might give us better insights into what other statistics go into player potential.