Predicting the Potential of players in FIFA 2019

Chandan Arora

May 18, 2019

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 rumoured transfer to the club in picture. Before spending millions on a player, the clubs analyse 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.

   - **Problem**

     The aim of this project is to predict the potential of a player based on the data present in the dataset. The dataset contains the details of players, their nationality, and other attributes such as dribbling, acceleration, stamina, shot-accuracy, etc.

   - **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 from this link. This dataset contains the players' details with attributes that would be useful in predicting the potential of the player.

   - **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 These null entries have been replaced by the mean of the attribute to remove any discrepancy.

   - **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:
     1. Age
     2. International Reputation
     3. Stamina
     4. Strength
     5. Aggression
     6. Composure
     7. Ball Control
     8. Dribbling
     9. Acceleration

ID	Name	Age	Overall	Potential	International Reputation	Stamina	Strength	Aggression	Composure	BallControl	Dribbling	Acceleration
158023	L. Messi	31	94	94	5.0	72.0	59.0	48.0	96.0	96.0	97.0	91.0
20801	Cristiano Ronaldo	33	94	94	5.0	88.0	79.0	63.0	95.0	94.0	88.0	89.0
190871	Neymar Jr	26	92	93	5.0	81.0	49.0	56.0	94.0	95.0	96.0	94.0
193080	De Gea	27	91	93	4.0	43.0	64.0	38.0	68.0	42.0	18.0	57.0
192985	K. De Bruyne	27	91	92	4.0	90.0	75.0	76.0	88.0	91.0	86.0	78.0

Fig. illustrating list of attributes used in the dataset.

3. 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.
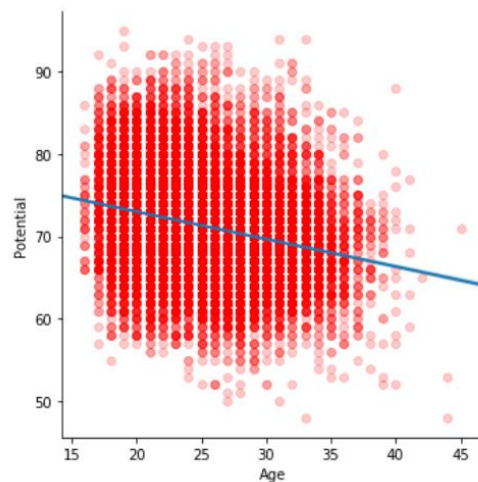
3.2 Potential & Stamina relationship

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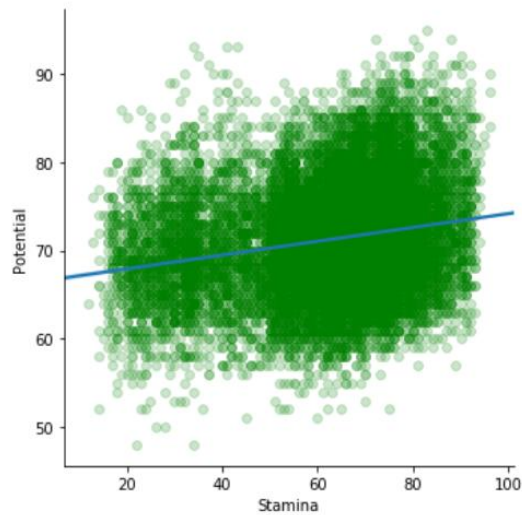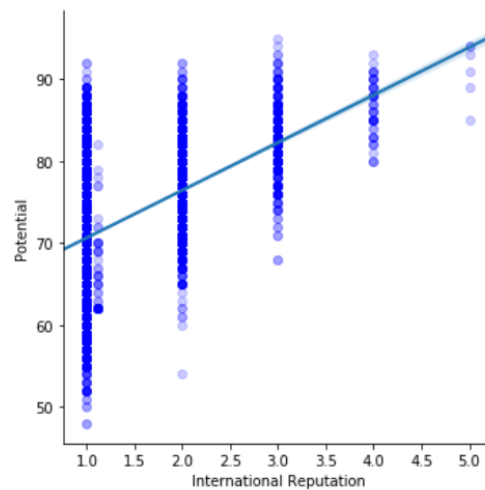

3.3 Potential & Age relationship

The following scatter plot shows the relationship between the target variable 'Potential' and 'Age' attribute.
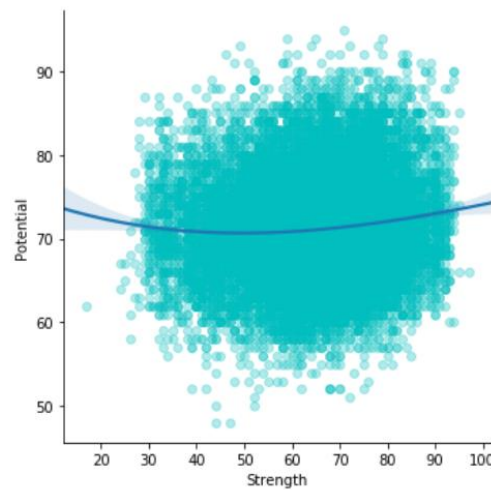
3.4 Potential & International Reputation relationship
The following scatter plot shows the relationship between the target variable 'Potential' and 'International Reputation' attribute.
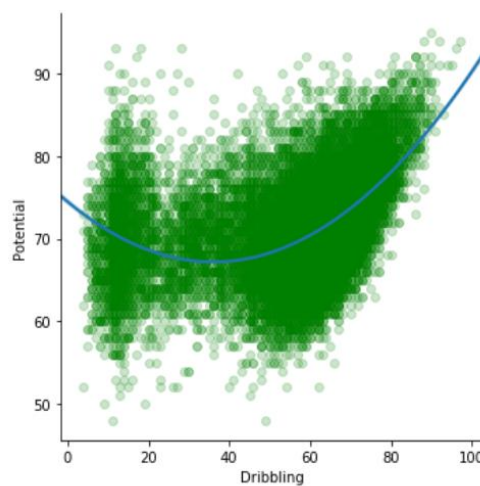

3.5 Potential & Strength relationship
The following scatter plot shows the relationship between the target variable 'Potential' and 'Strength' attribute.
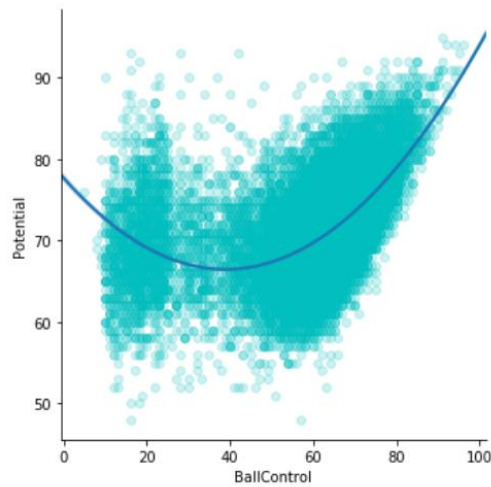

3.6 Potential & Dribbling relationship
The following scatter plot shows the relationship between the target variable 'Potential' and 'Dribbling' attribute.

3.7 Potential & Ball Control relationship
The following scatter plot shows the relationship between the target variable 'Potential' and 'Ball Control' attribute.

4. Predictive Modelling

I have used Regression model to predict the potential of a player based on other attributes present in the dataset. Linear Regression using one independent variable, Multiple Regression and Polynomial Regression have been used to predict the Potential and their accuracy has been compared.

1. **Linear Regression Model:** For Linear Regression, the attribute 'Strength' has been chosen as the independent variable, and 'Potential' as the target variable as defined above. The following joint plot shows the potential with respect to strength.
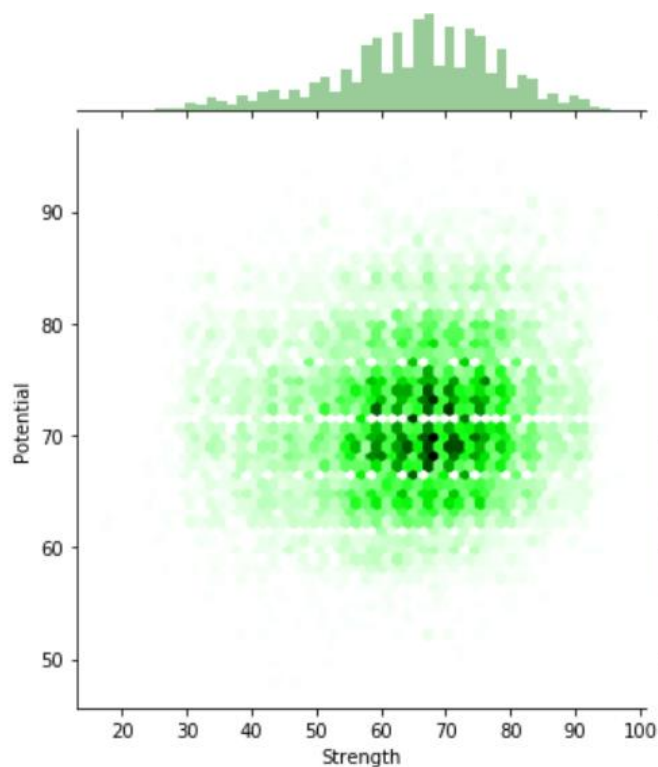
The following table shows the Intercept, Coefficient and the Mean Squared Error of the Linear Regression Model:

Intercept	69.00139693
Coefficient	0.03534763
Mean Squared Error	37.94121956223381

2. **Multiple Regression Model:** The following attributes have been chosen as the independent variables to find the target variable, i.e., Potential:

Independent Variables	Target Variable
1 International Reputation 2 Stamina 3 Strength 4 Aggression 5 Composure 6 Ball Control 7 Dribbling 8 Acceleration	Potential

The OLS Regression Summary is shown below:

OLS Regression Results

Dep. Variable:	Potential	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	1.641e+05
Date:	Mon, 20 May 2019	Prob (F-statistic):	0.00
Time:	09:54:14	Log-Likelihood:	-64517.
No. Observations:	18207	AIC:	1.291e+05
Df Residuals:	18199	BIC:	1.291e+05
Df Model:	8		
Covariance Type:	nonrobust		

The adjusted R-Squared means that 98.6% variables are explained in the dependent variable.

	coef	std err	t	P>|t|	[0.025	0.975]
International Reputation	6.2107	0.172	36.113	0.000	5.874	6.548
Stamina	-0.0730	0.007	-10.541	0.000	-0.087	-0.059
Strength	0.3862	0.005	74.870	0.000	0.376	0.396
Aggression	-0.1005	0.005	-18.584	0.000	-0.111	-0.090
Composure	0.4193	0.008	53.422	0.000	0.404	0.435
BallControl	0.1112	0.013	8.664	0.000	0.086	0.136
Dribbling	-0.2278	0.011	-20.840	0.000	-0.249	-0.206
Acceleration	0.4667	0.006	81.536	0.000	0.455	0.478

The t-values show that all the variables are significant and none of them can be ignored.

3. **Polynomial Regression:** In Polynomial Regression, I have used Strength as the independent variable to predict the potential of the players. The following scatter plot shows the relationship between Strength and Potential using Polynomial Regression of degree 3. The Mean Squared Error obtained from Polynomial Regression is .
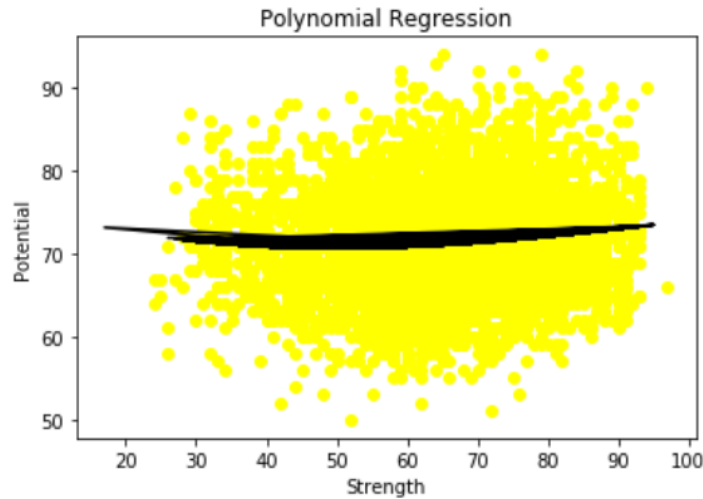
5. Conclusion

As the Mean Squared Error of Polynomial Regression (36.16809096502103) is less than that of the Linear Regression (37.94121956223381), therefore Polynomial Regression provides more accurate results.

6. Discussion

This model can be further analysed 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.