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

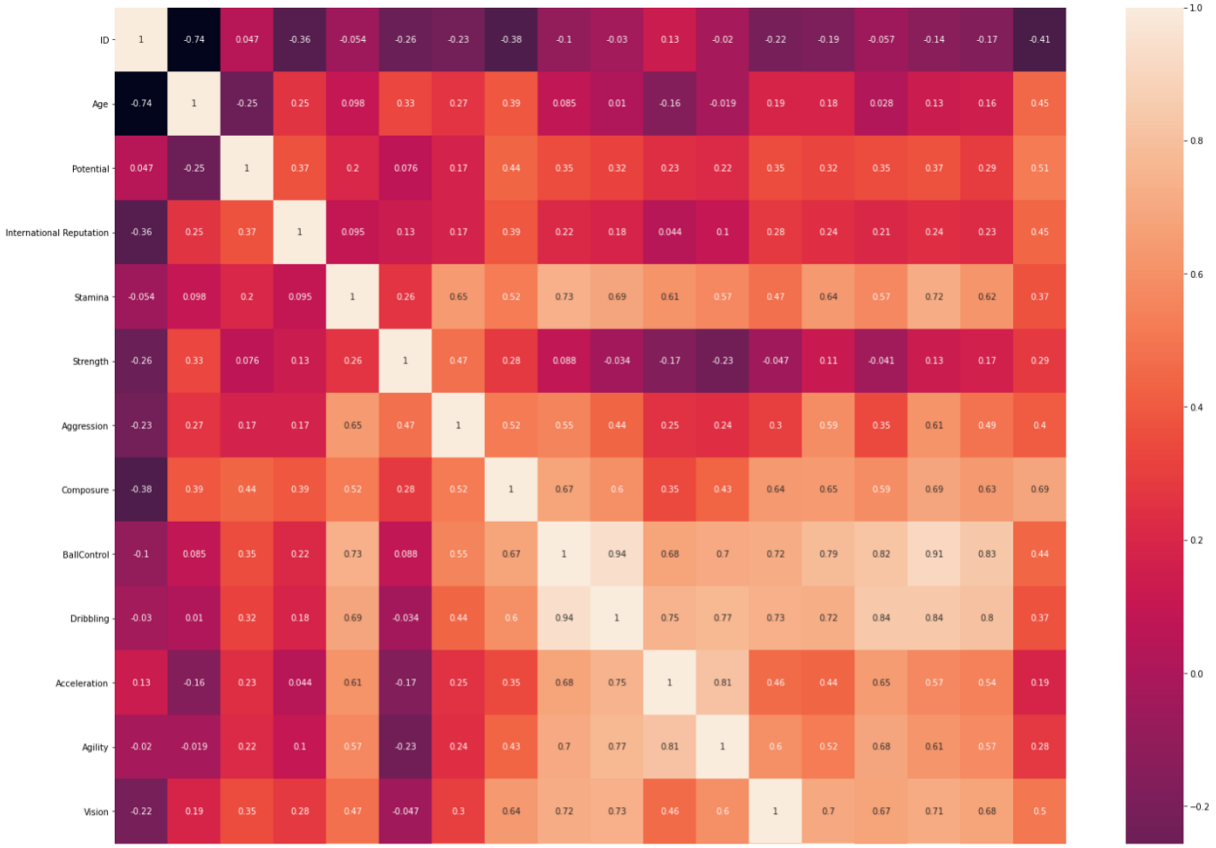
The dataset contains complete details of the players attributes such as age, preferred foot, weak foot, wages, skill moves, crossing, finishing, stamina, header accuracy, shot accuracy, etc. Some of the attributes such as stamina, strength, acceleration have a few null entries. These null entries have 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
10. Vision
11. Agility
12. Long Passing
13. Skill Moves
14. Short Passing
15. Shot Power
16. Reactions



3. Exploratory Data Analysis

3.1 Target Variable

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 Obtaining Relationships Between Target Variable and Features

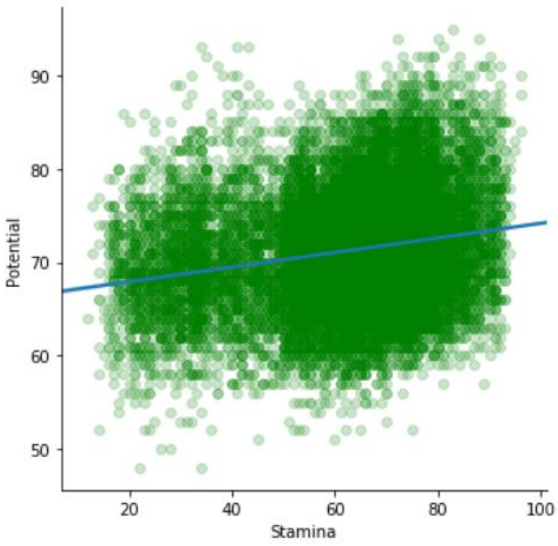
We need to obtain the relationship between the target variable and features selected and plot them in a graph to get better visualization of the dataset.

Understanding the interactions between multiple fields in the data set to make assumptions and predictions on the predictor variables that can help us to find the main target or to classify and cluster the data based on the selected variables.

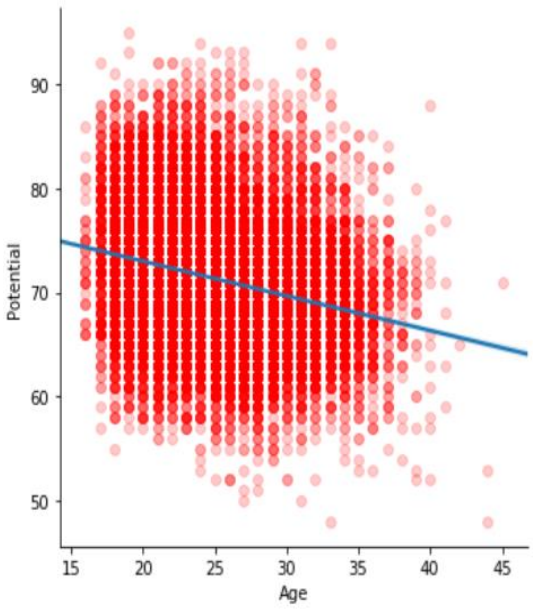
The data set consist of 185 features overall. However, not all the attributes would be required for regression techniques that are used here.

Initially all the relevant features that are required for making the predictions are picked based on the knowledge we possess. Much more relevant features are selected visualizing the relationship between the predictors and the response variables.

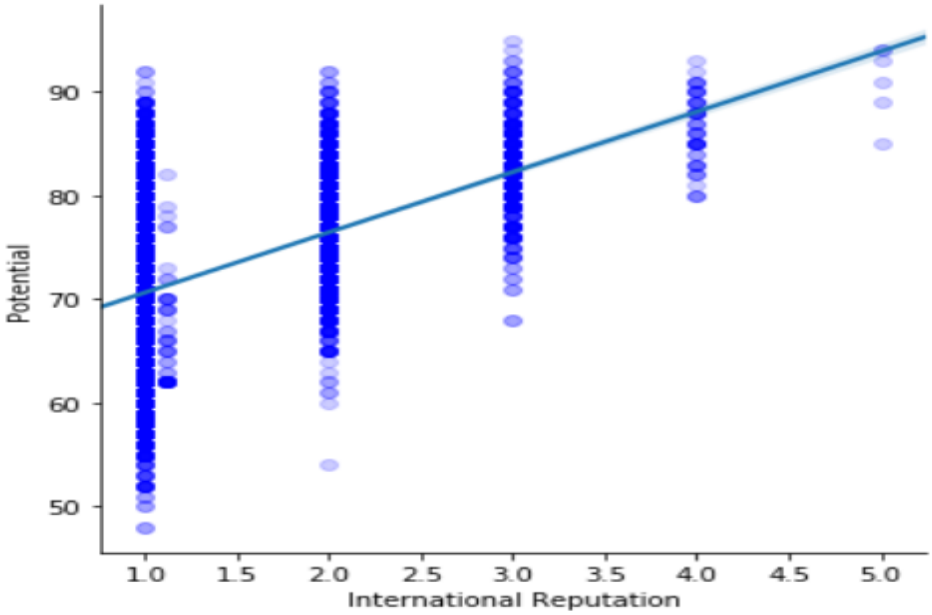
Potential & Stamina Relationship



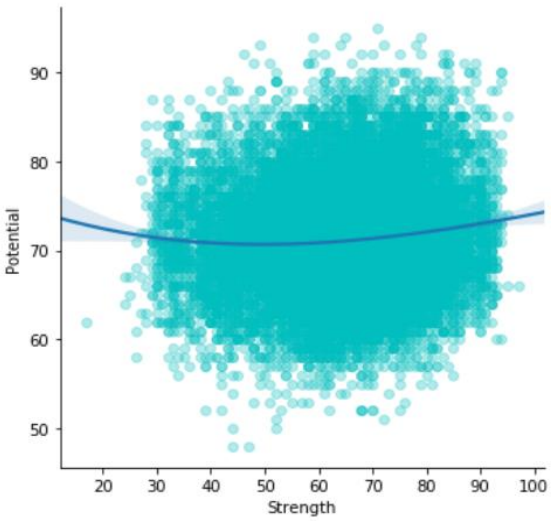
Potential & Age Relationship



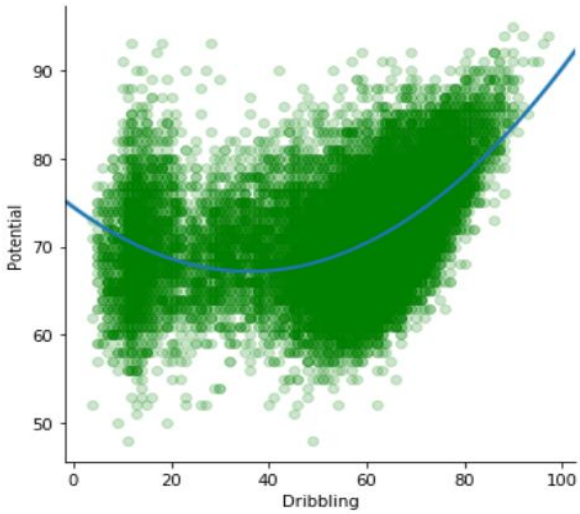
Potential & International Reputation Relationship



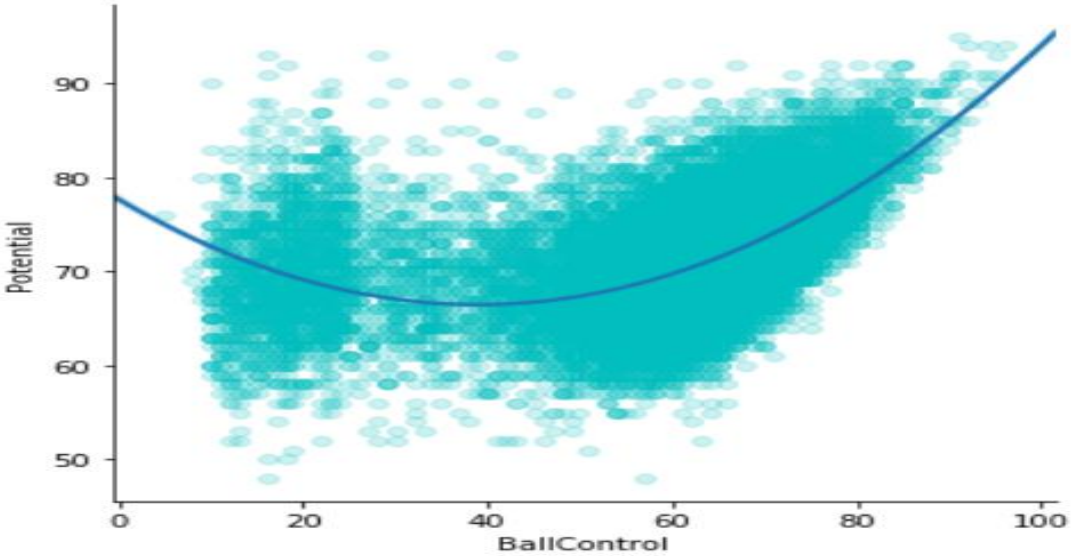
Potential & Strength Relationship



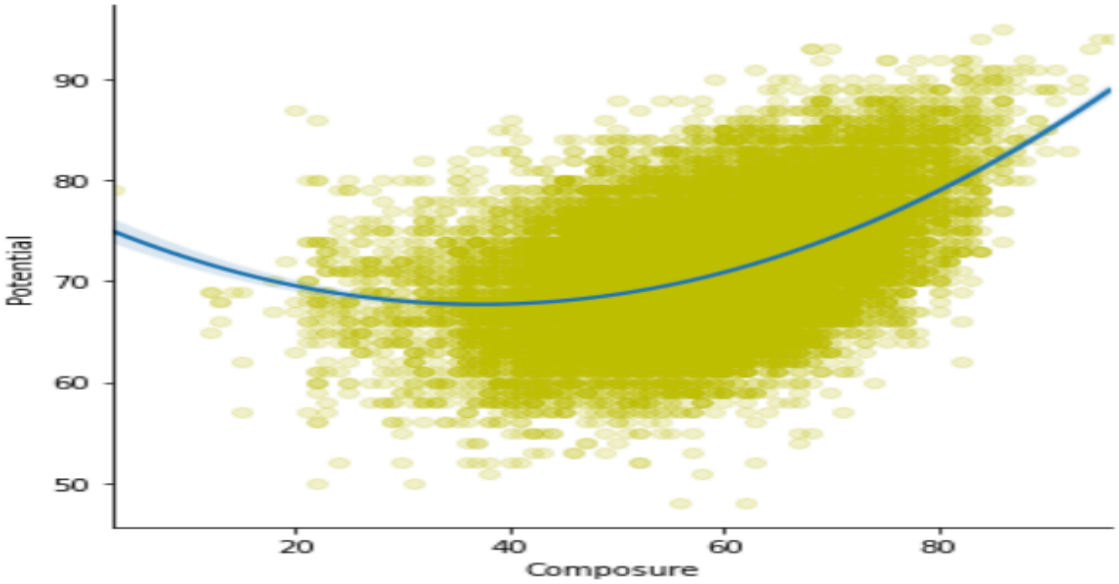
Potential & Dribbling Relationship



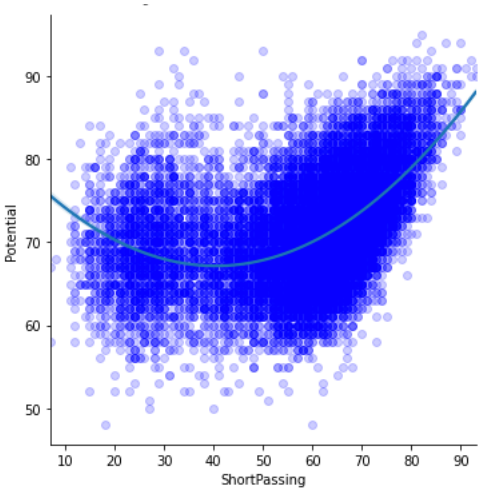
Potential & Ball Control Relationship



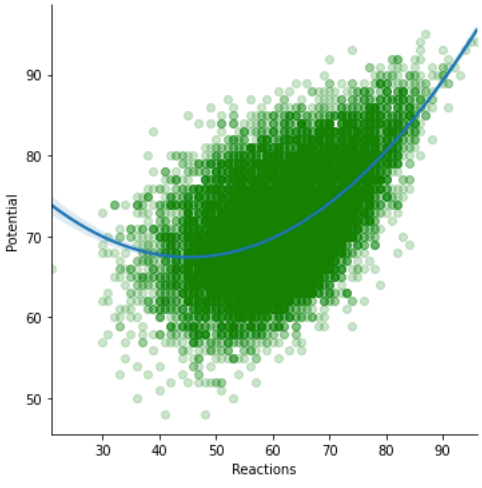
Potential & Composure Relationship



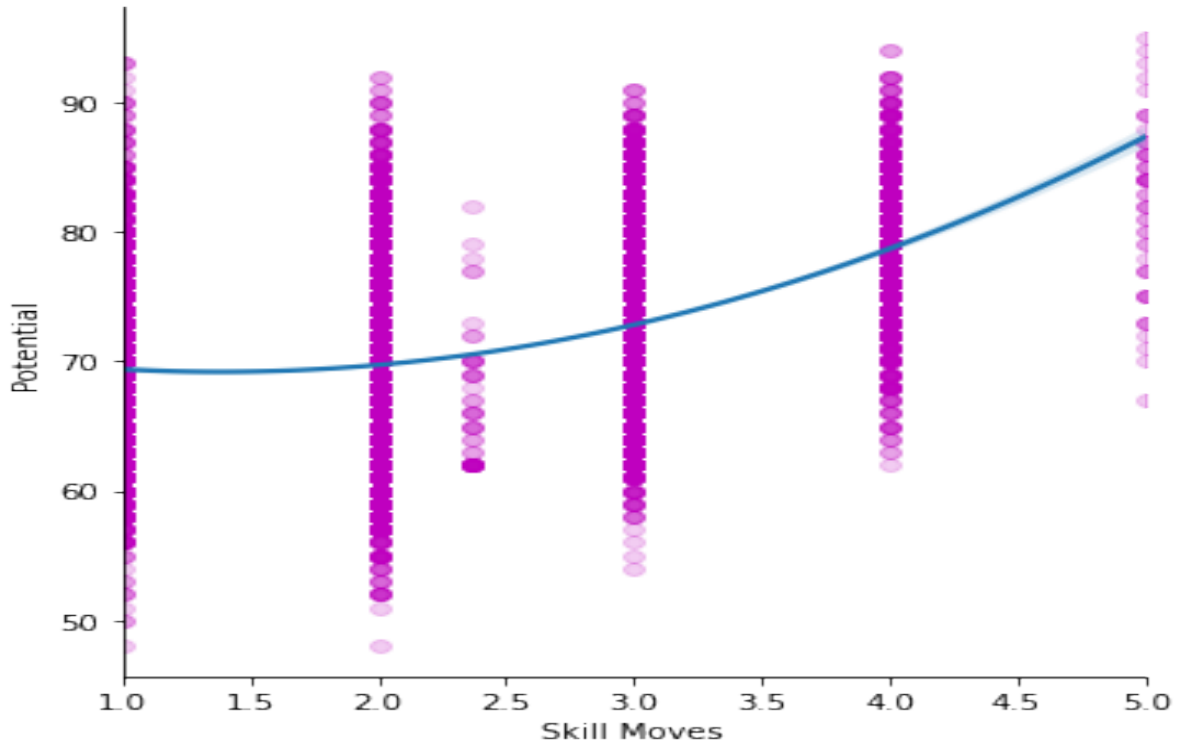
Potential & Short Passing Relationship



Potential & Reaction Relationship



Potential & Short Passing Relationship



The data set consist of 185 features overall. However, not all the attributes would be required for regression techniques that are used here.

Initially all the relevant features that are required for making the predictions are picked based on the knowledge we possess. Much more relevant features are selected visualizing the relationship between the predictors and the response variables

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, KNN, and XGboost to be used to predict the Potential and their Score Metrics have been compared.

4.1 Score metrics

For this problem we will use Mean Absolute Error and R-Squared as our metrics:

MAE:

$$score = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

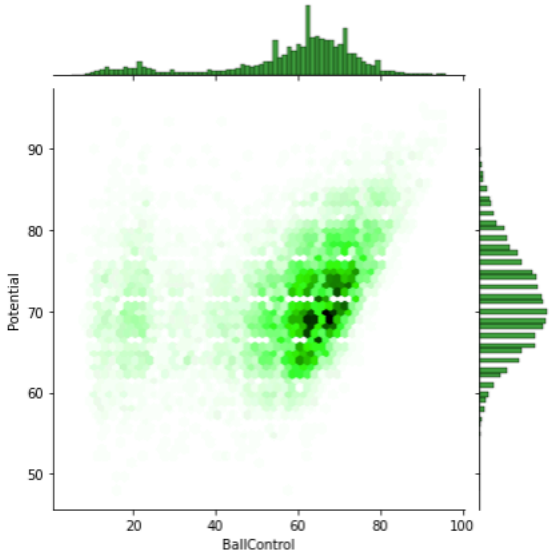
where *Yi* is the true value of *i-th* sample and -*yi* is the prediction for this sample. MAE Score signifies that average distance between prediction and true value; hence a lower score is better.

R-Squared:

$$R^2 = 1 - SS_{res} / SS_{tot}$$

where SS_{res} is the sum of squares of the residual errors and SS_{tot} is the total sum of the errors. A higher value of R2 is desirable as it indicates better results.

4.2. 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.



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

```
Training Set Accuracy: 0.12523800798660079
R-squared score:  0.12494625557275152
Mean Absolute Error: 4.606662359259951
```

Clearly, using just the highest correlated soccer statistic, BallControl, does not give our model a great R-Squared score or MAE 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.

4.2. Second Round of Modeling Using Different Regression Models

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

Independent Features	Target Feature
1. Age **2. International Reputation** **3. Stamina** **4. Strength** **5. Aggression** **6. Composure** **7. Ball Control** **8. Dribbling** **9. Acceleration** **10. Vision** **11. Agility** **12. Long Passing** **13. Skill Moves** **14. Short Passing** **15. Shot Power** **16. Reactions**	**Potential**

The OLS Regression Summary is shown below:

OLS Regression Results

Dep. Variable:	Potential	**R-squared (uncentered):**	0.991
Model:	OLS	**Adj. R-squared (uncentered):**	0.991
Method:	Least Squares	**F-statistic:**	1.268e+05
Date:	Fri, 23 Apr 2021	**Prob (F-statistic):**	0.00
Time:	14:25:51	**Log-Likelihood:**	-61174.
No. Observations:	18207	**AIC:**	1.224e+05
Df Residuals:	18192	**BIC:**	1.225e+05
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
International Reputation	2.4116	0.150	16.057	0.000	2.117	2.706
Stamina	-0.1024	0.006	-17.564	0.000	-0.114	-0.091
Strength	0.2785	0.005	57.924	0.000	0.269	0.288
Aggression	-0.0914	0.005	-19.093	0.000	-0.101	-0.082
Composure	0.0278	0.008	3.437	0.001	0.012	0.044
BallControl	-0.0389	0.013	-3.074	0.002	-0.064	-0.014
Dribbling	-0.1831	0.010	-18.502	0.000	-0.202	-0.164
Acceleration	0.2935	0.007	44.831	0.000	0.281	0.306
Agility	0.0929	0.007	13.583	0.000	0.079	0.106
Vision	0.0487	0.007	7.310	0.000	0.036	0.062
LongPassing	-0.0244	0.008	-2.991	0.003	-0.040	-0.008
Skill Moves	1.2453	0.133	9.373	0.000	0.985	1.506
ShortPassing	0.2360	0.012	19.472	0.000	0.212	0.260
ShotPower	-0.0747	0.006	-12.928	0.000	-0.086	-0.063
Reactions	0.5412	0.008	68.151	0.000	0.526	0.557

Omnibus:	794.156	**Durbin-Watson:**	1.728
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	909.137
Skew:	0.518	**Prob(JB):**	3.83e-198
Kurtosis:	3.352	**Cond. No.**	632.

The adjusted R-Squared means that 99.1% variables are explained in the dependent variable. The t-values show that all the variables are significant and none of them can be ignored.

Having added the extra independent features to the dataset, I fed the updated data into other regression models to see if I could improve the models' performance.

Here are the resulting test set R-Square and MAE numbers for each regression model used:

Multiple Regression Model

R-squared score: 0.3399158804514758
Mean Absolute Error: 3.950278271677799

Decision Tree Regression

R-squared score: -0.10190983498090689
Mean Absolute Error: 4.90782748278629

Random Forest Regressor

R-squared score: 0.4740246171650845
Mean Absolute Error: 3.4572610449733987

K-Nearest-Neighbors Regressor

R-squared score: 0.41914550734928224
Mean Absolute Error: 3.6298050521691376

XGBoost Regressor

R-squared score: 0.4576709257851318
Mean Absolute Error: 3.537189547535091

We see an improvement in R-Square and MAE score with some of our model regression models compared to our basic linear regression model.
The Decision tree regressor model is giving us a negative R-Square score, this means that our prediction tends to be less accurate that the average value of the data set over time.

4.3. Third Round of Modeling Using GridSearchCV

I'm going to select 3 of the models previously used and see if hyper-parameterizing our selected models using GridSearchCV can optimize our results.

Below are the results with their GridSearchCV optimal parameters:

Decision Tree Regressor

```
random_state=42,
criterion = 'mse',
max_depth = 8,
max_leaf_nodes = 100,
min_samples_leaf = 100,
min_samples_split = 1
```

R-squared score: 0.39279966768852415
Mean Absolute Error: 3.7532416506401964

Random Forest Regressor

```
random_state=42,
max_features='log2',
n_estimators= 500,
max_depth = None,
bootstrap = True,
criterion='mse'
```

R-squared score: 0.48036518344974855
Mean Absolute Error: 3.443018320263961

XGBoost Regressor

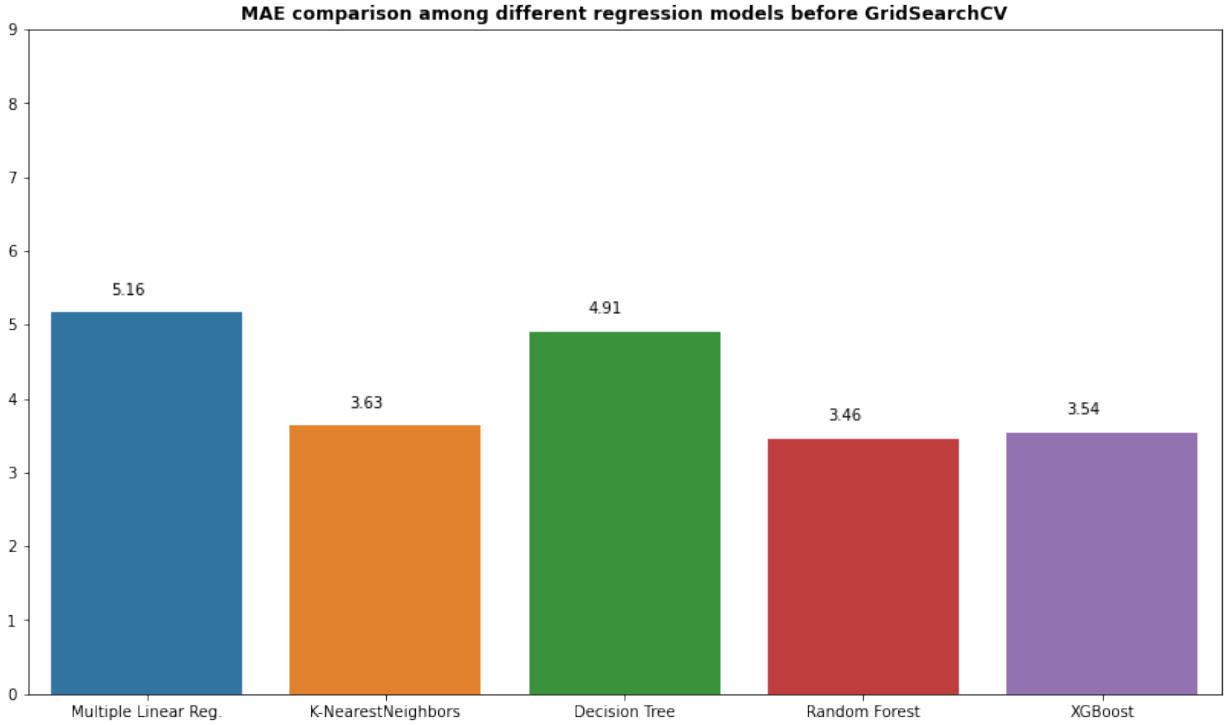
```
random_state=42,
colsample_bytree = 0.7,
learning_rate = 0.05,
max_depth = 5,
min_child_weight = 4,
n_estimators = 200,
nthread = 4,
objective = 'reg:linear',
silent = 1,
subsample = 0.7
```

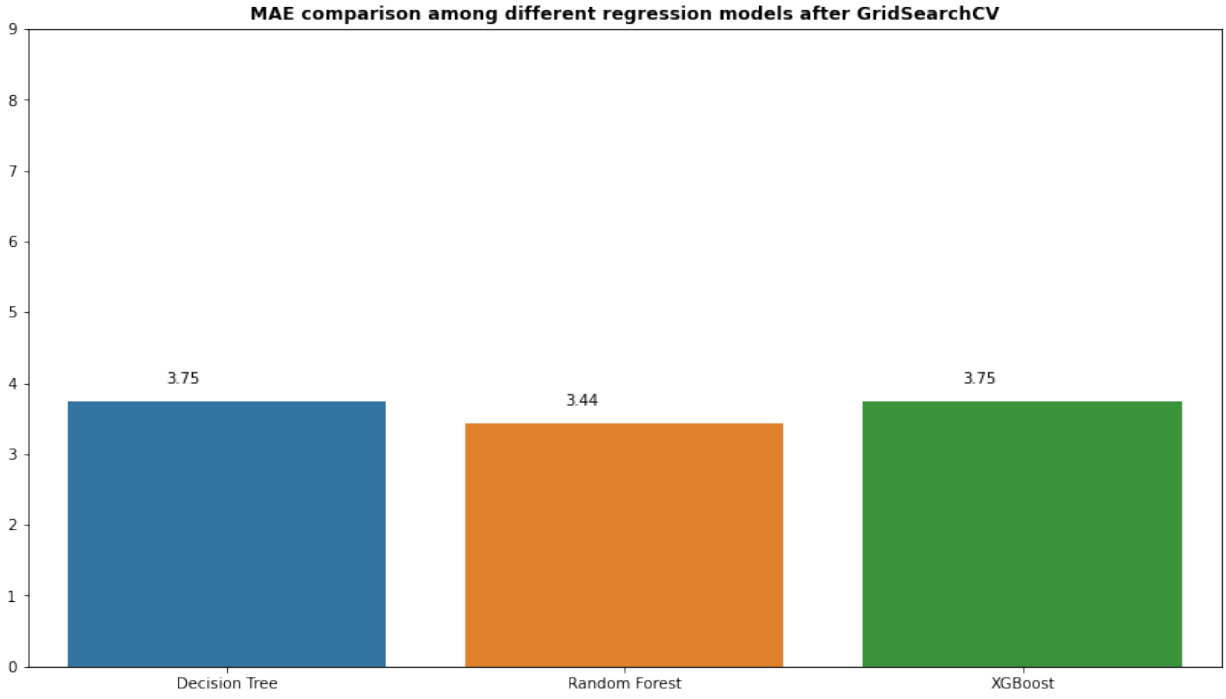
R-squared score: 0.39279966768852415
Mean Absolute Error: 3.7532416506401964

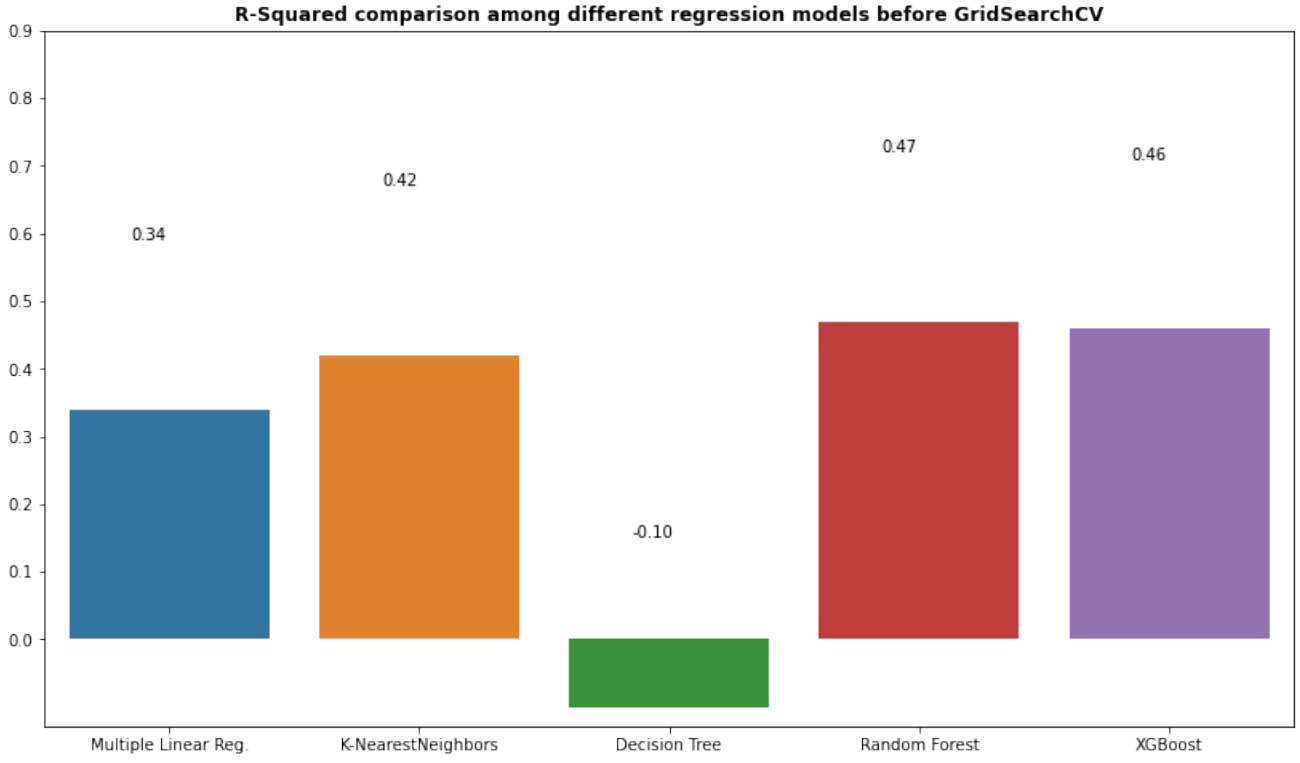
After running GridSearchCV on our selected models, we see a great improvement on the R-Square and MAE scores for our Decision Tree Regressor. R-Squared is no longer negative! XGBoost Regressor did slightly worse but that could be because of our chosen parameters. Overall, using GridSearchCV did improve our overall model.

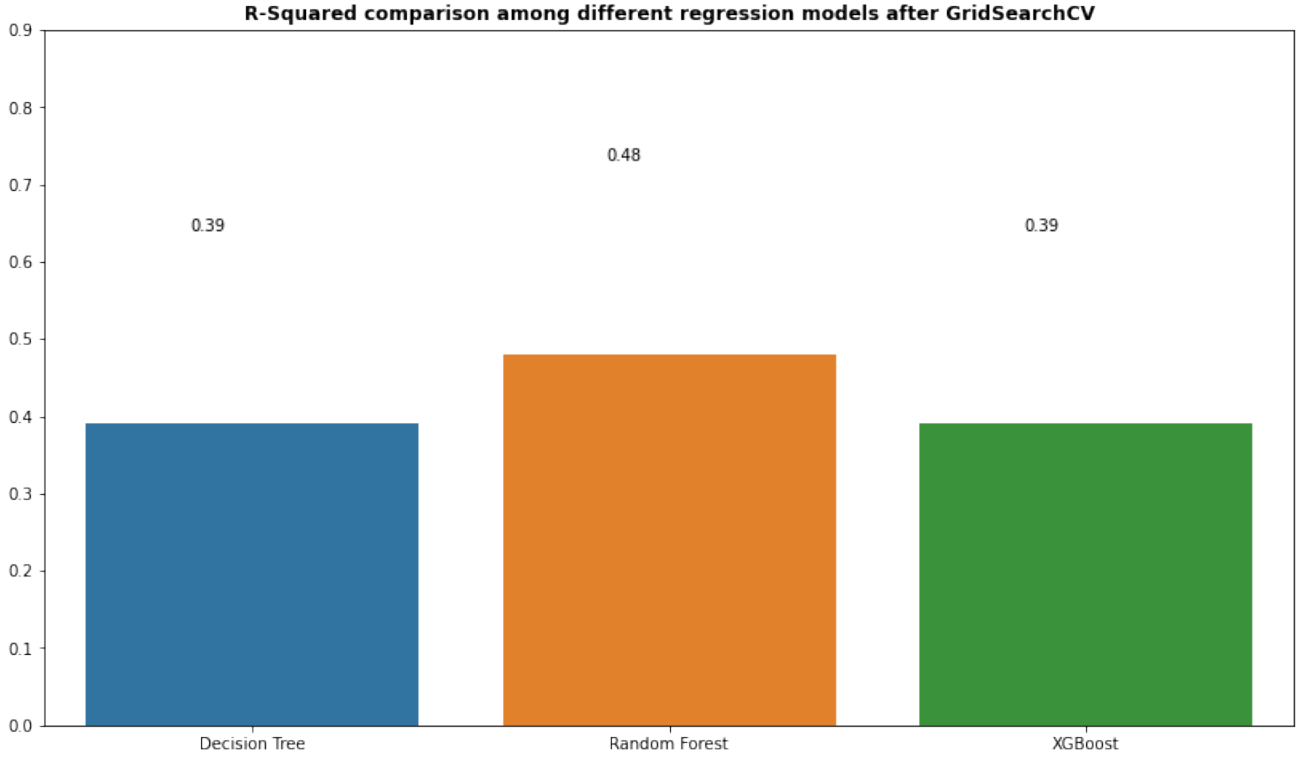
5. Conclusion

After looking at all our models, I would conclude that using either the Random Forest Regressor model model would be the best in predicting Player Potential. These models have the highest R-Squared score and lowest MAE Score out of all the models we tested. We purposely did not measure the models' accuracy because that metric isn't necessarily important here.






R-Squared comparison among different regression models before GridSearchCV
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.34
0.42
-0.10
0.47
0.46
Multiple Linear Reg.
K-NearestNeighbors
Decision Tree
Random Forest
XGBoost


R-Squared comparison among different regression models after GridSearchCV
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.39
0.48
0.39
Decision Tree
Random Forest
XGBoost

Feature Importance

Random Forest Coefficients for 10 Most important Features:

	Features	Coefs
5	BallControl	0.318354
14	Reactions	0.129700
7	Acceleration	0.053501
2	Strength	0.053081
1	Stamina	0.051193
13	ShotPower	0.051040
9	Vision	0.050524
12	ShortPassing	0.049990
6	Dribbling	0.048662
3	Aggression	0.048527

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. This can lead to our model having high score metrics.