

A close-up photograph of a soccer ball with black and white hexagonal panels hitting a white goal net. The ball is positioned in the center-left of the frame, and the net's hexagonal mesh is visible on the right. The background is a blurred green field.

Predicting Potential of players in FIFA

1. Introduction



Every year, after the end of season, the transfer window allows different clubs to buy, sell or loan players to/from other clubs



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

1. Introduction



Every year, after the end of season, the transfer window allows different clubs to buy, sell or loan players to/from other clubs



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

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](https://www.kaggle.com)



The complete dataset was downloaded in 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

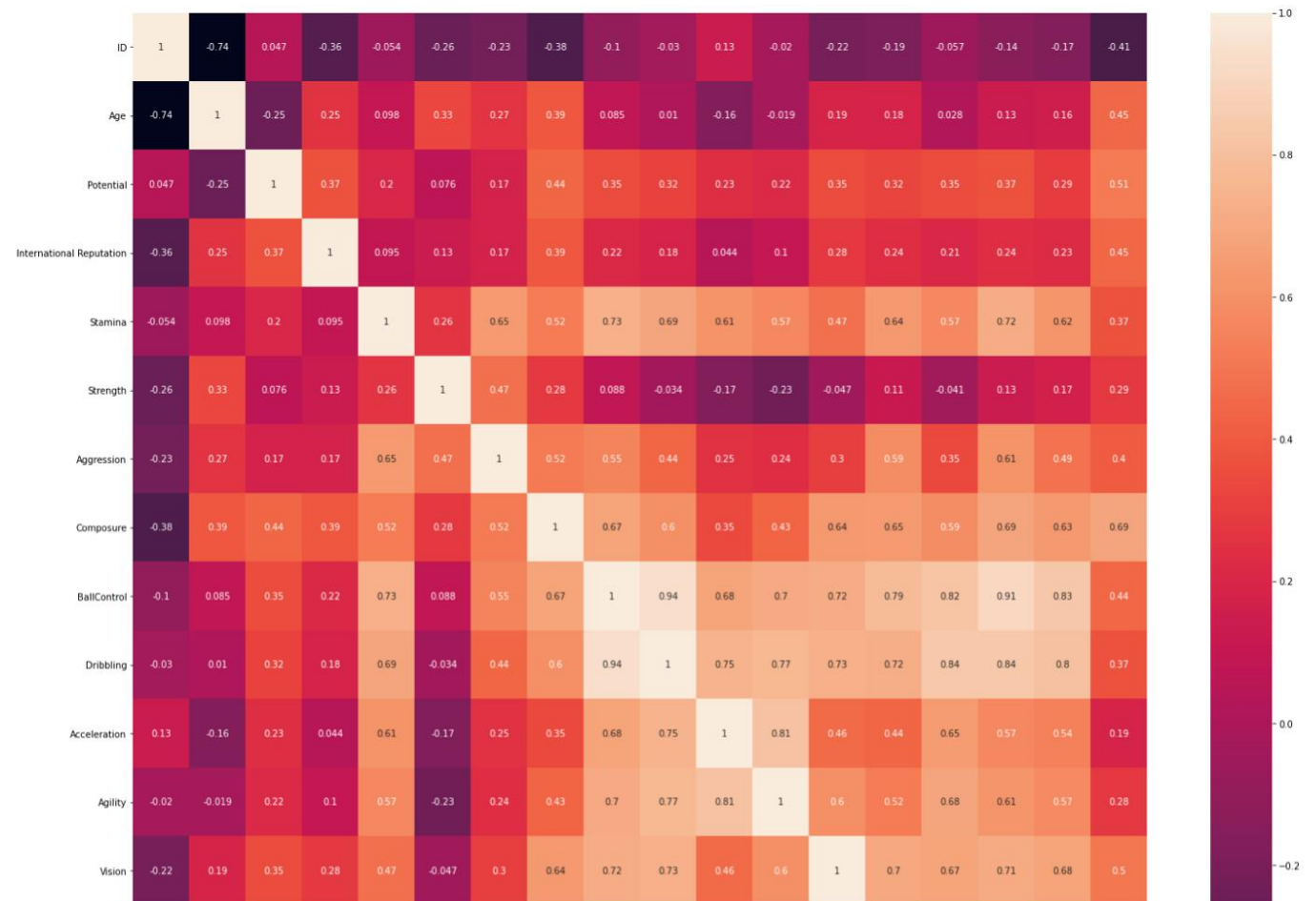
Some of the attributes such as stamina, strength, acceleration have a few null entries

These null entries have been replaced by the mean of the attribute to remove any discrepancy

2.3 Feature Selection

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

- Age
- International Reputation
- Stamina
- Strength
- Aggression
- Composure
- Ball Control
- Dribbling
- Acceleration
- Vision
- Agility
- Long Passing
- Skill Moves
- Short Passing
- Shot Power
- 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
- 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

Later, I implement further regression models such as Multiple Regression, Decision Tree Regression, Random Forest, KNN, XGBoost, to predict the Potential and their accuracy has been compared.

4.1 Score metrics

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

MAE/R-Squared



- MAE Score signifies that average distance between prediction and true value; hence a lower score is better
- A higher value of R^2 is desirable as it indicates better results

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

Second Round Model Scores

Multiple Regression Model	Decision Tree Regressor	Random Forest Regressor	KNN Regressor	XGBoost Regressor
R-squared score: 0.33991588045148	R-squared score: -0.10190983498091	R-squared score: 0.474024617165085	R-squared score: 0.419145507349282	R-squared score: 0.457670925785132
Mean Absolute Error: 3.95027827167779	Mean Absolute Error: 4.90782748278629	Mean Absolute Error: 3.45726104497339	Mean Absolute Error: 3.629805052169137	Mean Absolute Error: 3.537189547535091

4.3. Third Round of Modeling Using GridSearchCV

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

Decision Tree Regressor	Random Forest Regressor	XGBoost Regressor
R-squared score: 0.39279966768852	R-squared score: 0.480365183449749	R-squared score: 0.39279966768852
Mean Absolute Error: 3.753241650640196	Mean Absolute Error: 3.443018320263961	Mean Absolute Error: 3.753241650640196

5. Conclusion



After looking at all our models, I would conclude that using the Random Forest Regressor 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

A group of business professionals in an office setting. A man in a dark suit and striped tie is on the left, gesturing with his hand. A woman in a grey blazer is in the center, holding a smartphone. Another person is on the right, holding a white coffee cup. In the foreground, a tablet displays a document with circular diagrams. The scene is brightly lit, suggesting a window in the background.

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



This can lead to our model having high score metrics