FIFA-19 Player Value Prediction

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

Submitted by

Sreelekshmy S L, Nikhil Perumpulavil, Vidya G, Harsha, Arjun B K

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Feb 2022

List of Figures and Tables

Table of Contents

Abstract

This project focuses on player data from FIFA 19 which is developed by Electronic Arts for the 2019 edition of their FIFA game franchise in which you can play as many club and national soccer teams. The dataset used here is available on Kaggle. It has statistics on attributes such as preferred foot, body type, and skills in passing, shooting, and goalkeeping, that all go into a player's overall score.

The aim of our project is to predict the Value of a player with the given attributes. The dataset provides statistics of about 18000 players on about 60 different attributes. For each attribute, we have an integer from 1 to 99 that measures how good a player is at that attribute. These attributes are optimal indicators to determine the performance of a player at a particular playing position.

1. Problem Definition

1.1 Overview

FIFA 19 gives the users an option to play the game by selecting their desired team. It also gives the users an option to customize the team by hiring, upskilling (existing) or creating (new) players. While creating a dream team, the user might spend a major portion of their budget for signing few players which make it difficult to compete and advance in the game. Since the game automatically sets a standard value for each player, the users require a method to determine the value themselves. Our project aims to address these issues.

1.2 Problem Statement

Make a regression model to objectively determine the market value of a player based on the skill set they possess, thus allowing the user to form a team of excellent and undervalued players at minimal cost.

Introduction

Football (also known as **association football** or **soccer**) is a team sport played between two teams of 11 players each. It is played by 250 million players in over 200 countries, making it the world's most popular sport. Football has by far the most professional leagues in the world. The demand for football stars has been increasing dramatically in the past few decades, and the value of the football players is exceeding € 100 M.

Electronic Arts, the American video game company develop a series of football video games and release them annually, known as FIFA (also known as *FIFA Football* or *EA Sports FIFA)*. **FIFA 19** is the 26th installment in the FIFA series. This game provides proxies for actual player skill through their detailed rating system.

The FIFA 19 dataset which is available on Kaggle has statistics on attributes such as preferred foot, body type, and skills in passing, shooting, and goalkeeping, that all go into a player's overall score and value. A player's market value is an estimate of the amount with which a team could sell a player's contract to another team. In this project the FIFA 19 dataset is used to build a regression model that can predict a player's value. Also a simple website is created where the user can input the attributes in order to predict the player's value.

Data Understanding and EDA

We have done the EDA so as to analyze the dataset and summarize their main characteristics, with visual methods. EDA helped to gain a certain amount of familiarity with the data. Following steps are done in this section:

- Data Understanding
  - Attribute definition
  - Preview Data
  - Check total number of entries and column types
  - Check for null values

- EDA
  - Sorting nominal data by useful features
  - Univariate Analysis
    - Plot distribution of numeric data
    - Understanding the ordinal and categorical data
  - Bivariate analysis
    - Correlation of features
    - Comparison of features

Let us have a look at various insights we obtained from our data set:

Data Understanding

The dataset contains the statistics on the following attributes(35 out of 59):

Overall: General performance quality and value of the player.
Potential: Maximum Overall rating expected to be reached by a player in the top of his career rated between 1-99.
PreferredFoot: Right or Left.

WeakFoot: Represents how well a player uses his weak foot (e.g. left for righties) rated between 1 to 5.
WorkRate: Degree of the effort the player puts in terms of attack and defense, rated as low, medium and high.
Position: Position of the players on the pitch which determines their roles and responsibilities in the team.

Attributes that are rated between 1-99:

Crossing: Cross is a long-range pass from wings to center.
Finishing:Finishing in football refers to finishing an attack by scoring a goal.
HeadingAccuracy: Player's accuracy to pass or shoot by using his head.
ShortPassing: Player's accuracy for short passes.
LongPassing: Player's accuracy for long passes.
Dribbling: Dribbling is carrying the ball without losing while moving in one particular direction.
SprintSpeed: Speed rate of the player.
Acceleration: Shows how fast a player can reach his maximum sprint speed .
FKAccuracy: Player's accuracy to score free kick goals.
BallControl: Player's ability to control the ball.
Balance: Player's ability to remain steady while running, carrying and controlling the ball.
ShotPower: Player's strength level of shooting the ball.
Jumping: Player's jumping skill.
Penalties: Player's accuracy to score goals from penalty.
Strength: Physical strength of the player.
Agility: Gracefulness and quickness of the player while controlling the ball.
Reactions: Acting speed of the player to what happens in his environment.
Aggression: Aggression level of the player while pushing, pulling and tackling.
Positioning: Player's ability to place himself in the right position to receive the ball or score goals.
Vision: Player's mental awareness about the other players in the team for passing.
Volleys: Player's ability to perform volleys.
LongShots: Player's accuracy of shots from long distances.
Stamina: Player's ability to sustain his stamina level during the match. Players with lower stamina get tired fast.

Composure: Player’s ability to control his calmness and frustration during the match
Curve: Player’s ability to curve the ball while passing or shooting.
Interceptions: Player’s ability to intercept the ball while the opposite team’s players are passing. It is a defensive skill.
StandingTackle: Player’s ability to perform a tackle (take the ball from the opposite player) while standing. It is a defensive skill.
SlidingTackle: Player’s ability to perform a tackle by sliding.. It is a defensive skill.
Marking: Player’s ability to apply strategies to prevent opposing teams from taking the ball. It is a defensive skill.

While looking for the shape, we get that it has 18207 rows and 59 columns. Of the 59 columns, 13 are object type and 46 are numerical in the dataset. We described the dataset to get some insights like count, mean, min, max, quartiles. We found that the age range of players is between 16 and 45 with a median of 25. The overall score has a minimum value 46 and maximum value 94. Few columns like International Reputation, weak foot and skill moves seem to be discrete types with minimum value 1 and maximum 5. The columns from ‘Crossing’ to ‘GKReflexes’ in the order given are scores between 1 and 99. While checking for the null values, it is evident that 51 out of 59 columns have null values. Among them, most null values are in the ‘Loaned From’ column.

EDA

Firstly we sorted the data for a few nominal attributes and got the information that a large number of the players are from England followed by Germany and Spain. Juventus, Napoli, Inter, Real Madrid etc. are the clubs with high average overall scores. Juventus, Real Madrid and FC Barcelona are the clubs with high average value. L.Messi, Cristiano Ronaldo, Neymar Jr, De Gea are the players that topped the overall score. Also K.Mbappe, P.Dybala, G.Donnarumma are the players under the age of 25 with high potential.

Univariate Analysis:

When it comes to univariate analysis, from a histogram it is clear that most players are of age 20 to 30 with a right-skewed distribution. While ‘Overall’ is showing a normal distribution. Overall ranges from 46 to 94 with a median of 66 and have few outliers.

Value of players (Figure 1) has a minimum 10K and maximum 118.25M , with a median 700K and a lot of outliers.

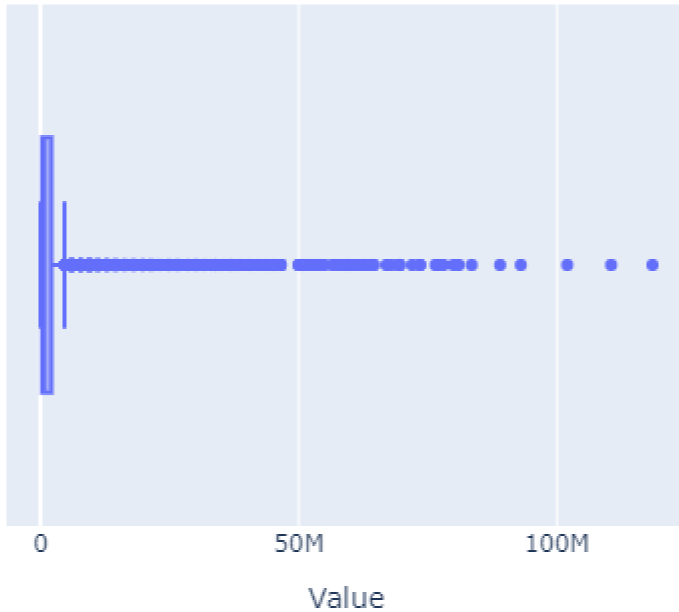

Figure 1: Distribution of Value

Then we have listed out the best players in different aspects such as 'Crossing', 'Finishing', 'Short passing', 'Volleys' etc. From a count plot showing the positions, it is evident that most players are strikers followed by goalkeepers and center backs. Since there are 27 positions we categorized them into four groups namely defenders,

Figure 2: Number of players by position category

midfielders, forwards and goalkeepers. And there are more midfielders than defenders and forwards (Figure 2). 'International Reputation', 'Weak Foot', 'Skill Moves' are the ordinal attributes ranging from 1 to 5. Most players have an ‘International reputation’ of 1, ‘Weak Foot’ of 3 and ‘Skill Moves’ of 2 ratings. Now the mostly found body types in the players are normal, lean and stocky. From a pie chart showing preferred foot, 76.81% of players prefer right foot and remaining left. Also only 9.11% of players are having their real face for their avatars in the game.

Bivariate Analysis:

While looking at Bivariate analysis, initially we have plotted a heatmap (Figure 3) to see the correlation between various features.

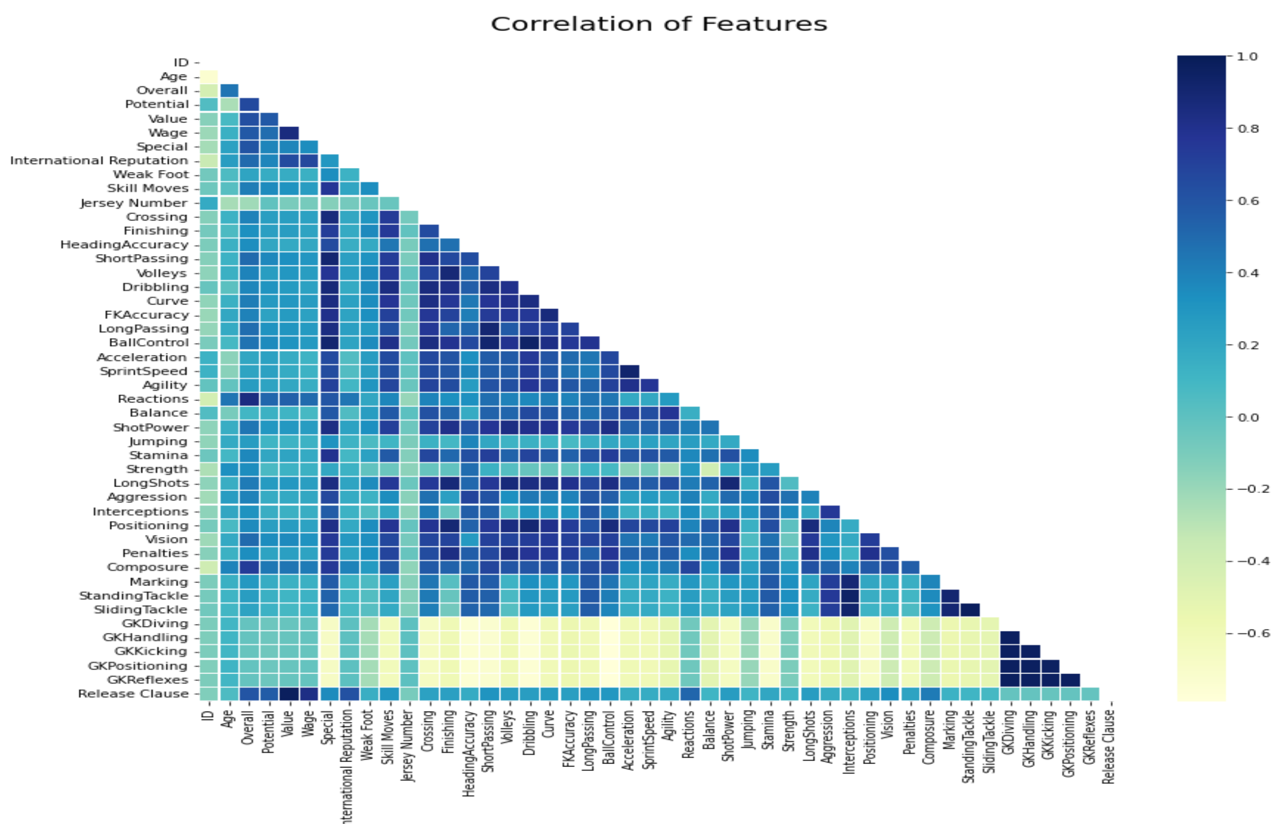

Figure 3: Correlation plot of features

We can see a pattern here for the columns ‘GKDiving’, ‘GKHandling’, ‘GKKicking’ and ‘GKReflexes’. These three columns are highly positively correlated to themselves, while they are highly negatively correlated to most of the other columns. ‘Release Clause’, ‘Wage’ and ‘International Reputation’ are highly positively correlated to ‘Value’. While plotting a scatterplot between ‘Value’ and these features (Figure 4) we can see that the ‘Release clause’ is almost double of the ‘Value’ of a player. Also for a few players ‘Release clause’ is set to be very low even after having good ‘Value’. After a certain ‘Value’ range, the ‘Wage’ of the players tends to scatter more. As ‘International reputation’ increases players have chances of getting higher ‘Values’ than previous one. But up to a rating of 4 most players share the same ‘Value’ range.

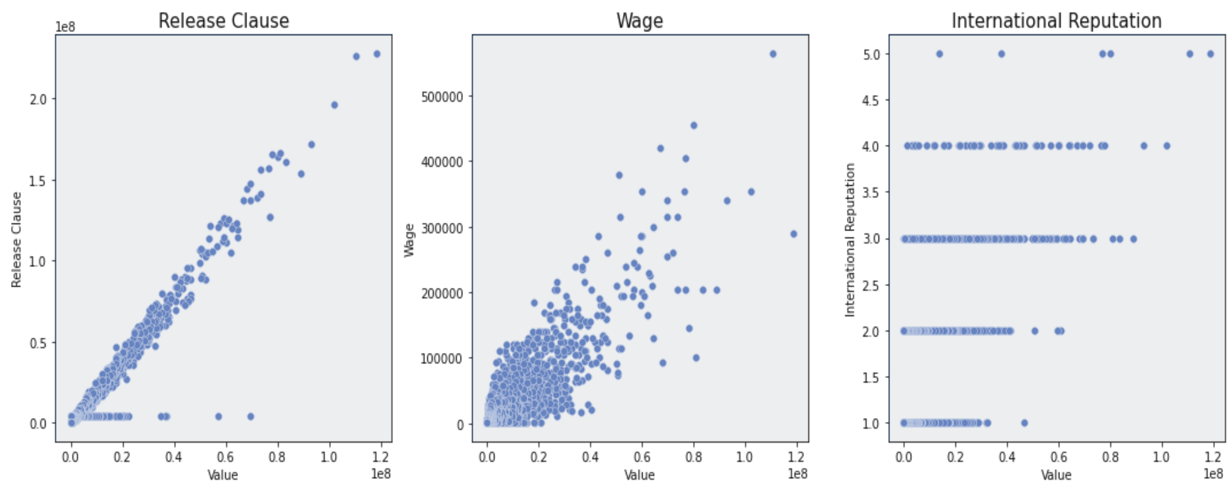

Figure 4: Scatterplots of ‘Release Clause’, ‘Wage’ and ‘International Reputation’ against ‘Value’

‘GKKicking’ , ‘GKHandling’ and ‘GKDiving’ are highly negatively correlated to ‘Value’. In both the plots (Figure 5) we can see two clusters one having ‘GKHandling’ and ‘GKKicking’ values between 0-20 and other having from 40 and above. The 0-20 cluster tends to have higher ‘Value’ than the other cluster.

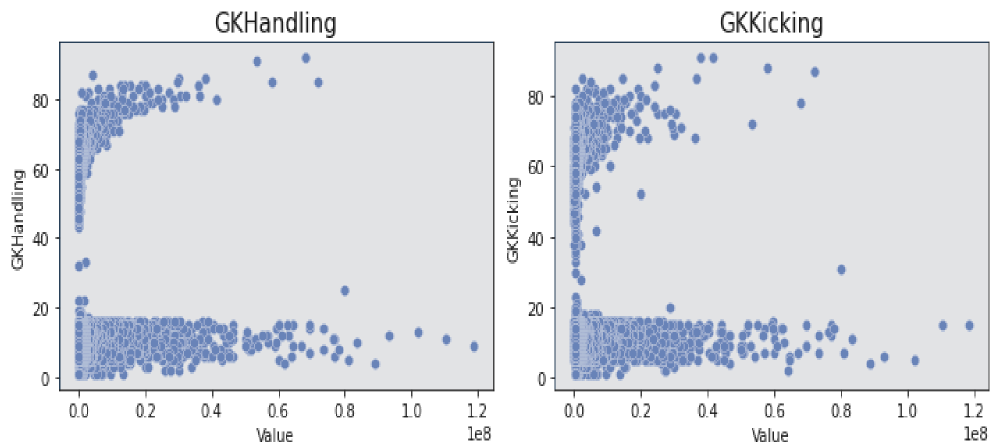

Figure 5: Scatterplots of ‘GKHandling’ and ‘GKKicking’ against ‘Value’

This could be because these features are specifically assigned to GoalKeepers. A box plot comparing ‘GKKicking’ of different positions explains this. ‘Reactions’, ‘Composure’, ‘Potential’, ‘Value’ are the columns highly positively correlated to 'Overall'. So we have plotted a scatter plot showing the relation between these features with ‘Overall’. In the ‘Overall’ vs ‘Potential’ plot, we could see that the potential is never below the overall score for a particular player. The value is almost constant until an overall score of 70 and then there is an exponential change in value as the overall score increases. We have plotted a scatterplot for ‘Overall’ vs ‘GKKicking’ and ‘GKDiving’ (highly negatively correlated features). Here also we could see two clusters, one is having a ‘GKDiving’/’GKKicking’ value in between 0 and 20, while the other has a value above 40. Since goalkeepers seem to be different from other positions we checked this with a scatterplot (Figure 6) for ‘Overall’ vs attributes that are not applicable to Goal keepers and found two clusters.

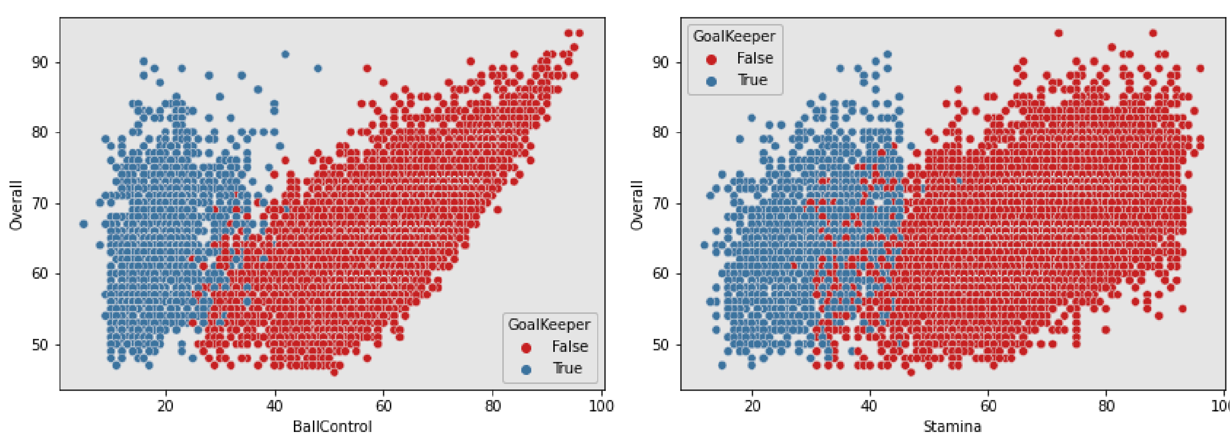

Figure 6: Plots showing 'BallControl' and 'Stamina' for 'Goal Keeper' and other positions

This might be because they're mainly confined to the goal and penalty areas, and exercise a limited range of motion compared to other players. This is also explained by a boxplot comparing BallControl of different positions. Scatterplot of 'Reactions' VS 'Composure' shows that they have a positive correlation and the 'Overall' score increases along with the increase in these two factors. Also we inferred that as the international reputation increases, wages also increase. We have used boxplots to compare 'Value' (Figure 7) as well as 'Overall' for different positions. Here we found that Forward and Midfielder positions have more 'Value' outliers.

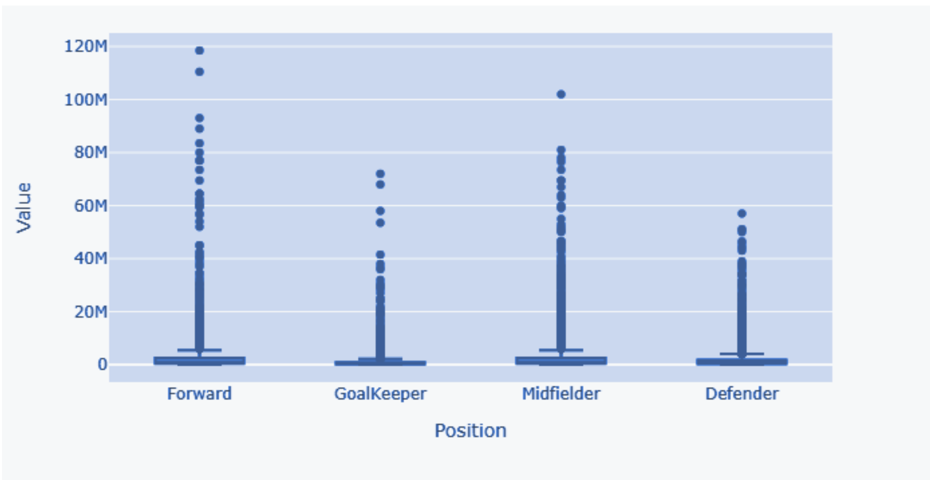

Figure 7: Distribution of Ball Control by position

Also the 'Overall' is slightly higher for Midfielder and slightly lower for Goalkeeper compared to other positions.

While comparing different features we get the following insights. From the boxplots comparing the 'Overall' and 'Potential' of players with their nationality (Figure 8), we get Brazil has a high average overall score, followed by Spain, France, and Argentina. In terms of potential, Spain seems to have overtaken Brazil, followed by Netherlands, France and Argentina.

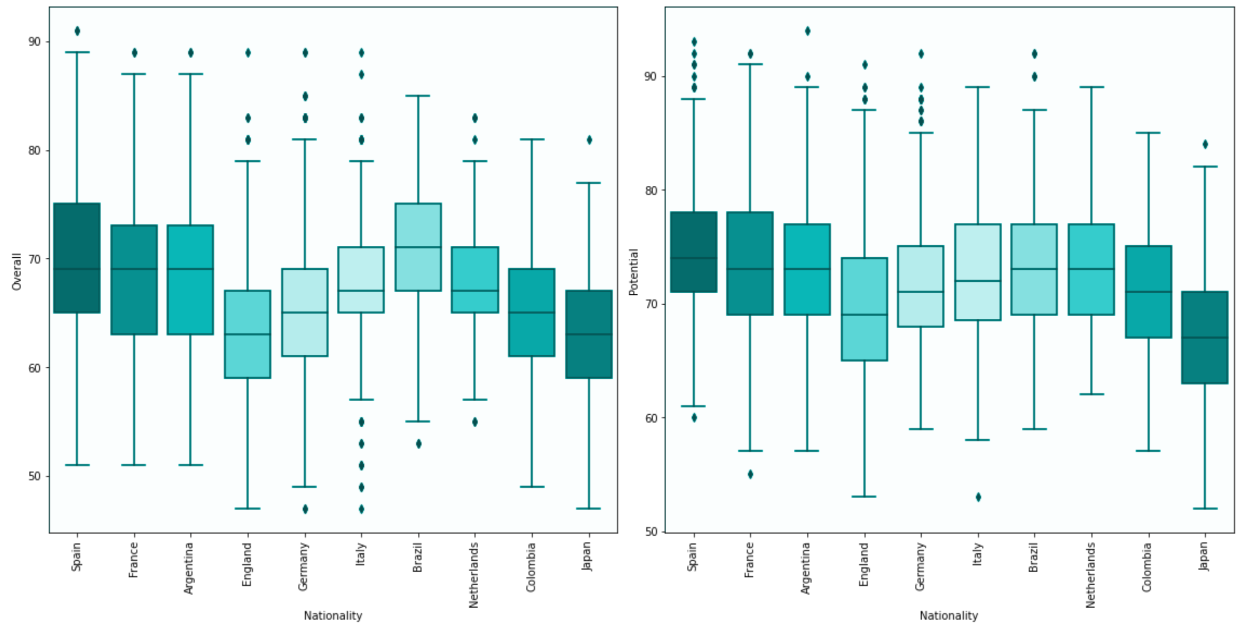

Figure 8: Overall VS Potential Score of top 10 nations

This indicates which teams will perform better in the coming few years. While comparing the ‘Overall’ and ‘Potential’ with the top 10 clubs using a boxplot (Figure 9), the median overall is seen to be highest for Juventis, followed by FC Bayern Munchen and Real Madrid . In terms of potential also, Juventis is highest while FC Barcelona seems to have overtaken Real Madrid and FC Bayern Munchen also Paris Saint-Germain has overtaken Inter and Napoli, indicating a presence of a lot of young talent in these clubs.

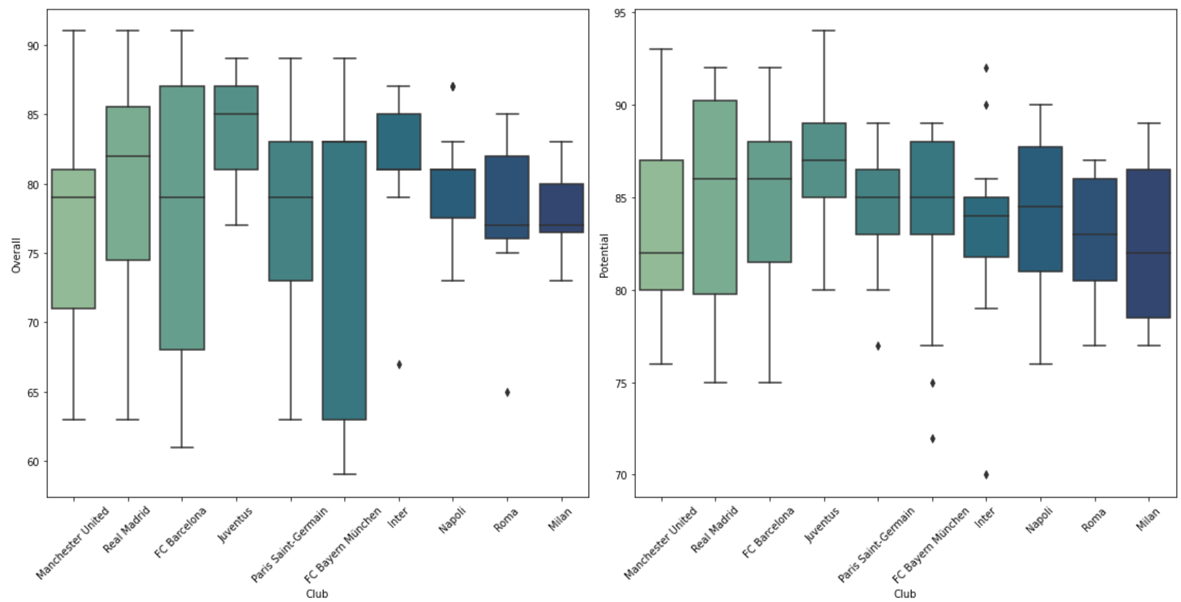

Figure 9: Overall VS Potential Score of top 10 clubs

When looking at the 'Overall' score, L.Messi and Cristiano Ronaldo scored the most followed by Neymar Jr and De Gea. But when it comes to Value of the player, Neymar Jr has the most value, L. Messi comes second and K.De Bruyne in third position (Figure 10).

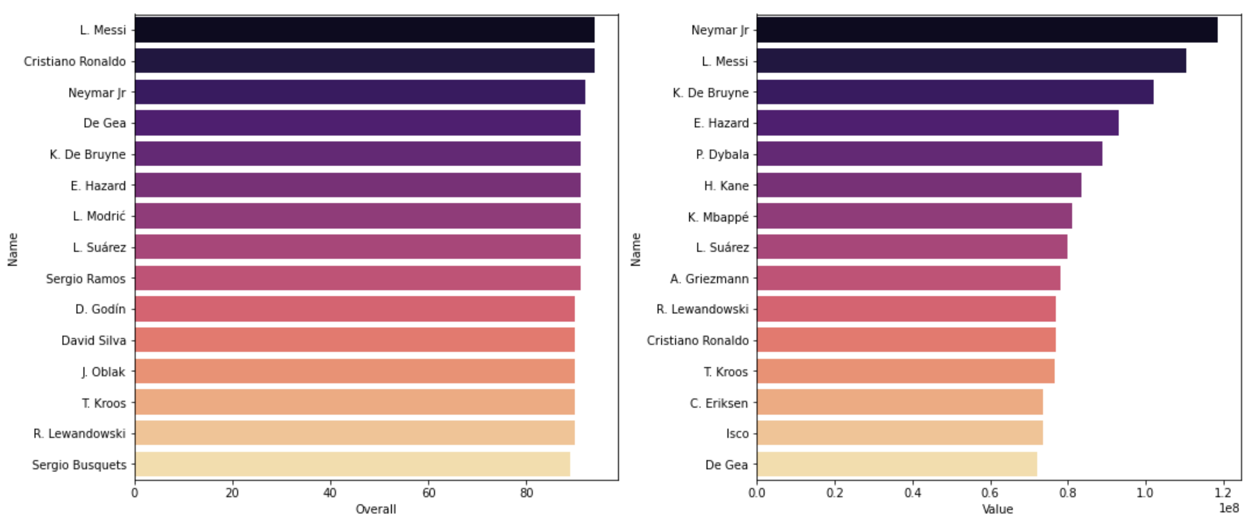

Figure 10: Overall Score VS Value of top players

We have sorted out the average of ‘Age’, ‘Value’ and ‘Wages’ for defenders, midfielders, forwards and goalkeepers. We found that defenders have a slightly high average age compared to others. The ‘Value’ and ‘Wage’ was more for forwards while it is very low for goalkeepers.

Data Preprocessing

Steps Followed :

- Missing Value handling
- Outlier detection
- Feature Engineering
- Encoding
- Standardization

As a first step we have dropped a few columns such as ‘ID’, ‘Jersey Number’ and ‘Real Face’ as these features don't affect a players’ ‘Value’.

Missing Value:

First we checked the columns with missing values and found that 49 out of 56 features have missing values in them. Among them, the ‘Loaned From’ column shares the highest percentage (93 %) so we dropped the column. Then in the ordinal columns like 'International Reputation' , 'Weak Foot' and 'Skill Moves' we filled the null values with the mode of the column. Next we considered the categorical columns. While checking the unique values in 'Preferred Foot','Work Rate','Body Type', we found that the ‘Body Type’ column has values other than the main three types, and some of them are famous player names. So we assigned them to the appropriate body types. Then we filled the missing values with the mode of the column. For the ‘Position’ column we checked the rows having null values and found that most of these columns are also missing individual score attributes. Hence a particular position cannot be assigned here. Also the ‘Club’ column cannot be assigned to the most frequent one. So we filled the null values of both of these columns with the string ‘N/A’ (Not Available). Checking the score columns we found all score columns have 48 null values and all these 48 rows are the same. So we decided to fill them with the median score value of their particular overall score group. In the ‘Height’ column the values were object type as it was in the form 5’7 representing 5 foot 7 inches. So we made a function to convert these values into cms. Similarly in ‘Weight’ column values have units added to them (eg : 100 lbs), here also we made a function to slice the numerical value alone. Both of these columns contain outliers so we filled the null values with median. ‘Joined ’ and ‘Contract Valid Until ’ columns were filled with the string ‘N/A’ for null values. From our EDA we learned that the columns ‘Value’ and ‘Wage’ are highly related to ‘International Reputation’. So we filled their null values with the median under a particular reputation rate. Atlast we rechecked the null values and all of them were removed.

Outlier detection:

During the EDA as well as missing value treatment portions, we found that a lot of columns have outliers in them. Hence we re-checked it for a few specific columns . First we checked ‘Age’, ‘Overall’, and ‘Potential’(Figure 11). There are few players around the age 40 to 45, also some players have an overall score and potential score below 50 and above 90. Since these values are still in the permitted range we cannot drop or replace them.

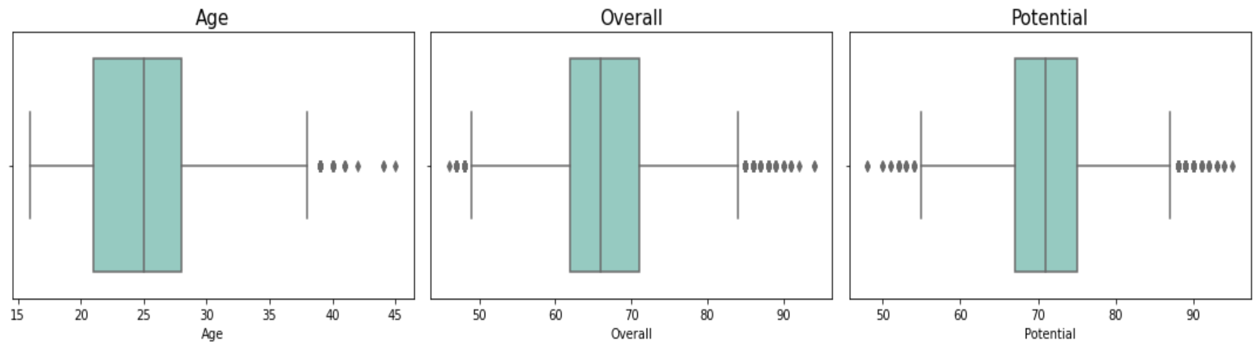

Figure 11: Boxplots of ‘Age’, ‘Overall’ and ‘Potential’

In the ‘Value’ and ‘Wage’ column there are plenty of outliers(Figure 12). But we know these depend on a players reputation and overall score we cannot handle them.

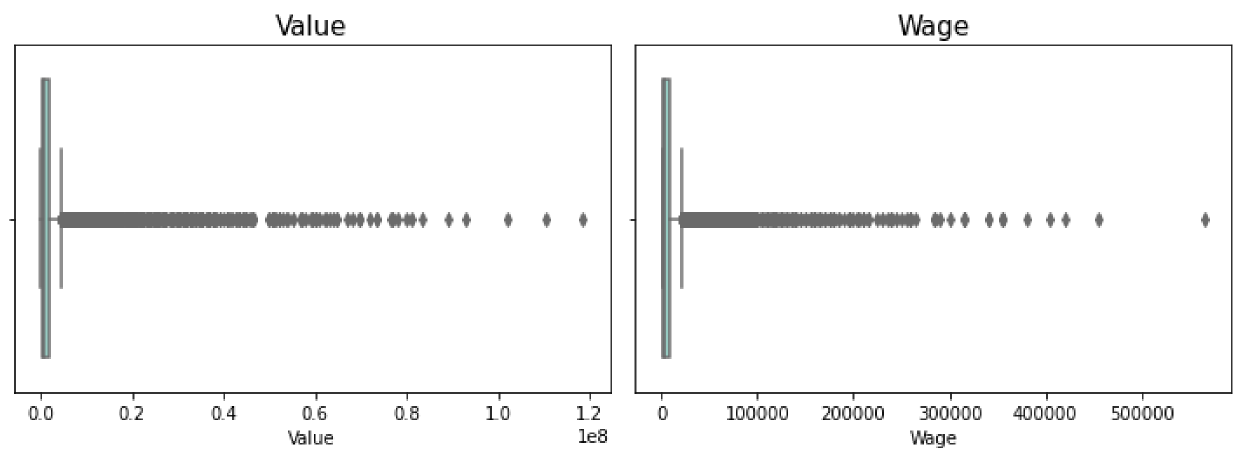

Figure 12: Boxplots of ‘Value’ and ‘Wage’

While checking the ‘Height’ column we could find that one of the outliers has a value near 0, which is not possible to occur. So we replaced it with the median of height. In the score columns, those columns representing power features of a player show outliers in the range 0-30 (Figure 13).

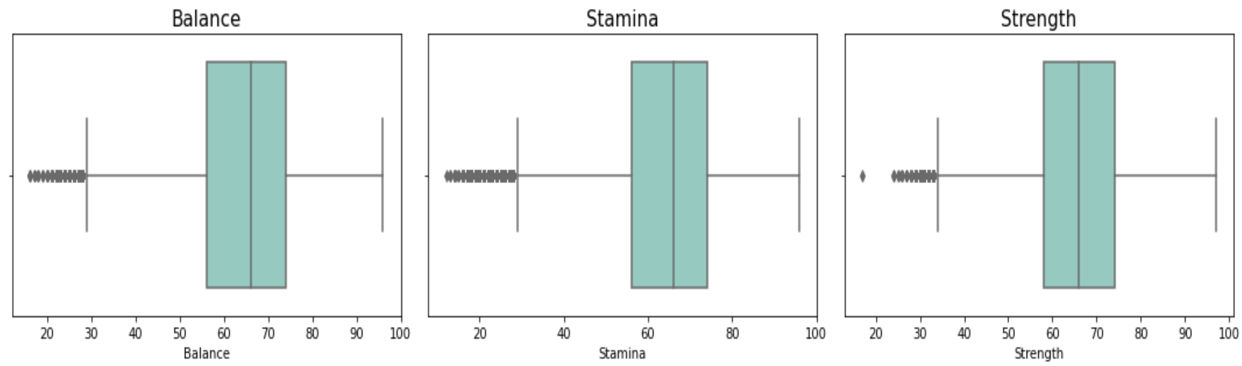

Figure 13: Boxplots of ‘Balance’, ‘Stamina’ and ‘Strength’

As mentioned in EDA, these features are low for Goalkeepers, and that is the reason for these outliers. Similarly for features specific to Goalkeepers, the outliers are found in the range 40 to 100 (Figure 14).

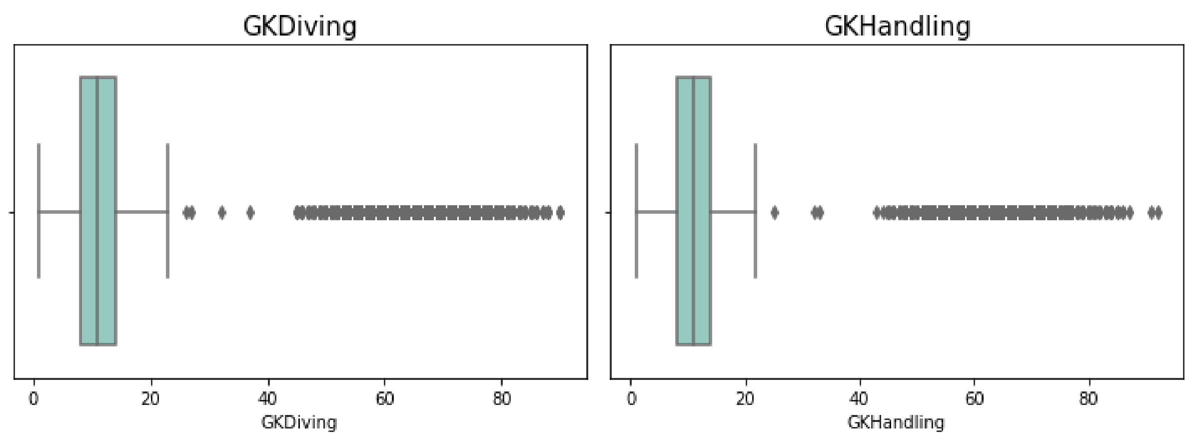

Figure 14: Boxplots of ‘GKDiving’ and ‘GKHandling’

All these cannot be replaced or removed as they have a reason for their existence and changing them can affect our prediction model.

Feature Engineering :

We have around 27 positions, all of which make it really hard to classify. We first categorized them into four main groups as we have done for the EDA and a new column ‘Position_group’ is added to the dataset showing the position group of each player. These groups are named as Forward, Midfielder , Defender and Goalkeeper. We then calculated the scores specific to each position group for every player by taking the average score based on the below classification.

- Forward Attributes: 'BallControl', 'Positioning', 'Reactions','Composure', 'Finishing', 'ShotPower','ShortPassing', 'Dribbling', 'Volleys', 'LongShots'
- Midfielder Attributes: 'BallControl', 'Reactions','ShortPassing','Composure', 'Vision','Dribbling', 'LongShots', 'LongPassing'
- Defender Attributes: 'StandingTackle', 'Interceptions', 'SlidingTackle', 'Reactions', 'Marking', 'Composure', 'ShortPassing'
- Goalkeeper Attributes: 'GKReflexes', 'GKDiving', 'GKPositioning', 'GKHandling', 'Reactions', 'GKKicking'

  This will give the users an idea about which position a player is best suited at. We may also drop the above feature columns for our further analysis. Further, in case of creating new players in the game, the users will have a better control on the features required for the player.

Encoding:

Next we considered the object type columns. ‘Name’ and ‘Nationality’ columns can be dropped. Contract valid column is encoded with 1 for the years 2022 and 2023 all others as 0. The ‘Joined’ column is split into month and year columns for better understanding. The ‘Work Rate’ column is split into ‘Attack rate’ and ‘Defence rate’ by the ‘/’ symbol in it. They are now having values low, medium and high. So we encoded them as 1, 2 and 3 respectively and fixed the data type as numeric. Now the columns ‘Position_group’ and ’Body type’ are one hot encoded, while ‘Preferred foot’ is label encoded. After all these we rechecked the dimension of the dataset, the number of columns is now 64, also no rows were removed.

Feature Scaling:

Firstly we have splitted our target variable from attributes and assigned it to y. The columns except Value (our target variable), Join year and Club are assigned to X. Standard scaling, Normalization and Min-Max Scaling are done on X. Then we applied a linear regression model to each of them and compared the R square values. We found that Standard Scaling and min max scaling has no drastic effect on the model as the r_2 value is almost the same before and after. But Normalizing has decreased the r_2 value by a large amount. So we decided not to apply any scaling to our dataset.

Feature reduction:

Heatmaps of the feature columns are plotted to find highly correlated features. ‘Special’ has a high correlation with forward_score, midfielder_score and defender_score. So ‘Special’ is dropped. ‘Value’, ‘Wage’ and ‘International Reputation’ are highly correlated. Since these features are required for our further analysis and modeling, they are retained. Here the individual score columns like 'BallControl', 'Positioning', 'Reactions' etc (total 31) are also dropped since the newly created grouped score columns are enough to determine a player's performance in his position group.

As a last step we have rechecked our dataset and found that the number of rows remains the same (18207) as we haven’t dropped any player from the dataset. The number of columns has been reduced to 39 after dropping and adding a few. All of the columns now have numeric data types, which is a must for modeling.

Modeling and Fine Tuning

The aim of our project is to predict the ‘Value’ of a player based on other attributes. Hence we need a regression model. We will first try different machine learning models such as Linear Regression, Decision Tree etc along with an ensemble model and a boosting model. The evaluation metrics ‘Mean Squared Error’ (or MSE) and ‘R square’ (or r2_score) will be calculated for each model on the test data. Comparing these values we will find the best fitting model for our dataset. Before modeling we first removed the ‘Release Clause' column as it is dependent on ‘Value’ and not the other way around.

Modeling:

Since our target column ‘Value’ consists of large digit numbers we have converted this into a two digit float value using the formula log(1+x), where x is the individual value. Next we have split the data into train and test sets, here we took the test size as 20% of total data. Now in the train set we have 14565 rows and in the test set we have 3642 rows also we have 37 columns on both sets. Then we applied the following regression models on our data : LinearRegression, RidgeCV, LassoCV, Decision Tree, RandomForestRegressor and XGBRegressor. For each of these models we first imported the necessary libraries. Most of them are in the sklearn library and one is from xgboost. Then we created an instance of the model, and we fitted it with the train data set, after that we predicted the y value for the test data and stored it in a variable. All of these models were used with their default parameters. Next we calculated their MSE and r2_scores, stored them and displayed them individually.

Model	MSE	R-square
Linear Regression	0.073516	0.961127
Ridge Regression	0.073471	0.961150
Lasso Regression	1.303145	0.310932
Decision Tree	0.045531	0.975924
Random Forest	0.022197	0.988263
XGBoost Regressor	0.024573	0.987007

Table 1: Mean Square Error and R-square for each model

We then compared these two values by making a table (Table 1). And found that all the models except LassoCV have the MSE less than 1 and r2_score greater than 0.95. Hence most models are performing well. And among them the Random Forest Regressor has the lowest MSE(0.022) and highest r2_score (0.988). So we have found one of the best fitting models for our dataset and it is the Random Forest Regressor.

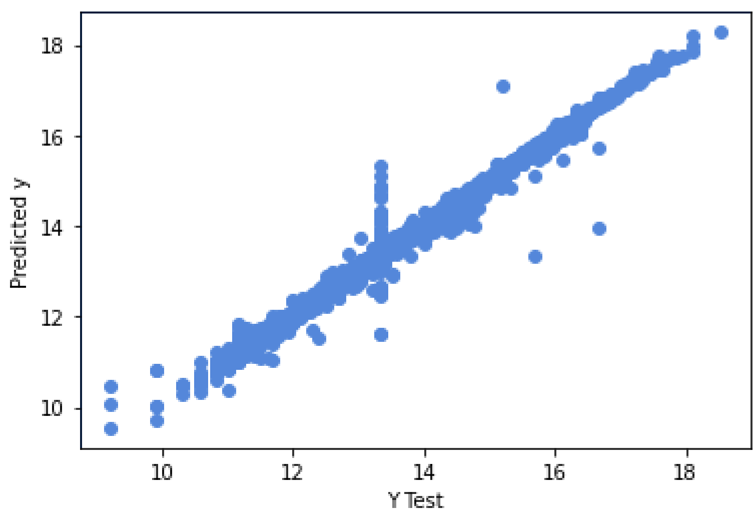

Figure 15: Plot of predicted 'Value' VS test set 'Value'

Then we plotted a graph between the actual y value as well as the predicted y value for the test data (Figure 15). The scatterplot shows a thick straight line with few outliers indicating most of the predicted values were close to the actual value. We then also checked the actual and predicted values of the first 5 rows in test data (Table 2).

ID	Actual	Predicted
5848	1200000.0	1168401.0
14634	270000.0	262036.0
11807	400000.0	412793.0
5826	775000.0	777450.0
1989	9000000.0	8089588.0

Table 2: Actual and Predicted ‘Value’

While checking the feature importance, we get that ‘Overall’, ‘Potential’, ‘Age’, ‘Wage’,‘forward_score’, ‘penalties’ are the top contributors in the list. And ‘Overall’ score has a very large feature importance score (86.63).

Fine Tuning:

In this step we tried to improve our Random Forest Regressor model using hyper parameter tuning on features like ‘n_estimators’, ‘max_features’, ‘min_samples_split’ and ‘bootstrap’. First we used Randomized search technique as it is less time consuming and these were the best parameters chosen: n_estimators=100, min_samples_split=8, max_features='auto', bootstrap=True. While checking the metrics MSE and r2_score we found that these changes did improve our model slightly. Now we used GridSearch with the same parameters for more efficient choice and we got : min_samples_split=2 and all others were similar as previous. Now when we checked the evaluation metrics it has increased the r2_score and decreased the MSE value, thus improving our model.

Web Hosting

For web hosting we have taken the following features from our dataset and used it to predict the player Value:

'Age', 'Overall', 'Potential', 'Preferred Foot', 'forward_score', 'midfielder_score', 'defender_score', 'goalkeeper_score', 'International Reputation', 'Weak Foot', 'Skill Moves'

We have used flask for web hosting and here is the website link that was created using pythonanywhere.

https://tinyurl.com/2p952xv4

Result

The Random Forest regression model is finalized to predict the value of a player with the given attributes. After necessary hyper parameter tuning our model has a MSE value of 0.0218 and a R- squared value of 0.9884, which is a good score.

Conclusion

FIFA 19 automatically sets a standard price for each player in the game. It is, therefore, necessary for a user to objectively determine the value of the player to form a team of excellent and undervalued players at minimal cost. For this player market value regression task, we had better results with a Random Forest Regressor. The user can now provide the skill sets of the required player in the website and the model will predict the player’s value. Then the user can determine whether to choose that particular player or create a new player with the same skill sets.

References

1. Kaggle dataset: https://www.kaggle.com/karangadiya/fifa19
2. FIFA player details: https://sofifa.com/
3. Wikipedia : https://en.wikipedia.org/wiki/FIFA_19
4. Data visualization using plotly:https://www.geeksforgeeks.org/using-plotly-for-interactive-data-visualization-in-python/
5. Evaluation metrics for regression: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/
6. Random Forest Regression : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
7. Hyper parameter tuning for random forest: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74