**FIFA 20**



**Selecting FIFA 20 Best Squad**

**January 2021**

**Agenda**

1. **Problem Statement**
2. **Project Objective**
3. **Python Resources**
4. **Data Description and Location**
5. **Methodology**
   1. **Data Preprocessing**
   2. **Correlation Analysis**
   3. **Feature Selection**
   4. **Model Selection**
   5. **Prediction**
6. **Key Visualizations and Findings**
7. **Project Recommendation**

**1. Problem Statement**

Player needs to build the best squad possible for a FIFA 20 tournament, but the total value of the 11 players chosen cannot exceed the amount of **€*275M***. Additionally, the team should have one Goal Keeper, 2 or 3 Defenders, 1 or 2 Attacking Defenders, 2 or 3 Midfielders, 1 or 2 Attacking Midfielders, and 2 or 3 Strikers.

**2. Project Objective**

Develop, design, and implement a solution that accurately predicts and classifies players by tiers, to identify player performance similarities and player´s best values.

**3. Python Resources**

Python contains different libraries and options to perform the project analysis. For this particular exercise we used the following resources:

|  |  |
| --- | --- |
| **Resources** | **Details** |
| **Main libraries** | Pandas  Numpy  Scipy  Math - sqrt |
| **Data Visualization** | Matplotlip.pyplot  Seaborn |
| **Metrics** | Sklearn.metrics – mean\_squared\_error  Sklearn.metrics – r2\_score  Sklearn.metrics - accuracy\_score |
| **Modeling** | Sklearn.model\_selection – cross\_val\_score  Sklearn.model\_selection – train\_test\_split  Sklearn.feature\_selection – RFE  Sklearn.model\_selection - GridSearchCV |
| **Classifiers** | Sklearn.linear\_model – LogisticRegression  Sklearn.neighbors – KneighborsClassifier  Sklearn.ensemble – RandomForestClassifier  Sklearn.svm - SVC |

**4. Data Description and Location**

The data chosen is located in the Kaggle Open Datasets in the following url:

<https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&gclid=CjwKCAiA57D_BRAZEiwAZcfCxaOHm8Qx6QzO5WLZsLoxnOoxF8o9g9vvK28xB-A1agtUsIBjngewgxoCwqUQAvD_BwE>

Two datasets were chosen for the project, *“FIFA 19 complete player dataset”,* and *“FIFA 20 complete player dataset”*. The first one was used to train, test, and create the model, and the second was used to implement it and identify a recommendation.

Both datasets had the same information structure, and the following variables:

|  |  |
| --- | --- |
| **Observation** | **Description** |
| Player characteristics | Data like name, age, height, weight, DOB, nationality, club, preferred and weak foot |
| Player details | Information such as positions, jersey number, value, wage, release clause, contract validity |
| Player ratings | Different measures from 1 to100 (higher number better) on areas such as pace, shooting, passing, dribbling, defense, physic, and goal keeping. |

**5. Methodology**

**5a. Data Preprocessing**

***Initial data review and cleanup***

Upon data upload, an initial review identified that the dataset is composed of 104 columns with 17.770 observations each. After this initial step, a preliminary data cleanup was done, eliminating repetitive and non-value columns such as player url, id, and player tags. This resulted in trimming the variables to just 60.

***Missing Values***

Data was checked for missing information, and identified 12 columns with incomplete data. The columns were:

|  |  |  |  |
| --- | --- | --- | --- |
| ***pace*** | ***shooting*** | ***passing*** | ***dribbling*** |
| ***defending*** | ***Physic*** | ***gk diving*** | ***gk handling*** |
| ***gk kicking*** | ***gk reflexes*** | ***gk speed*** | ***gk positioning*** |

For all “gk” categories data was completed with a 0 value, as the data came from players that were not goalies, and therefore should not be good or skilled at the position.

For all other categories, the missing data came from all goalies. Given that they are not usually good field players, data was completed by assigning the average lowest score of all the six main rating variables.

***Variable coding and discretization***

The “preferred foot” variable was coded as 1 for “Right”, and 2 for “Left”. Same was done for player positions and created 6 categories:

|  |  |  |
| --- | --- | --- |
| ***1 = Goalkeeper*** | ***2 = Defender*** | ***3 = Attacking Defender*** |
| ***4 = Midfielder*** | ***5 = Attacking Midfielder*** | ***6 = Striker/Forward*** |

The “overall rating” variable was discretize into a new column composed of 5 tiers. This new variable would be the dependent variable for the project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***1 = Excellent*** | ***2 = Above Avg*** | ***3 = Average*** | ***4 = Below Avg*** | ***5 = Poor*** |

The rationale behind this is that although there are differences between players ratings, you can group them into tiers or groups of similar skill and performance. This makes it easier to understand which players have similar output despite differences in their specific rating, aiding the selection of the best player and value.

As a last step, observations were assured to have the correct data type (int/float) and eliminated object variables used just for data visualization purposes (Short Name, Club, and Nationality).

**5b. Correlation Analysis**

A correlation analysis was done for the data set. It was identified that the only correlation with the dependent variable came from the “overall rating”. Also collinearity was found between many independent variables. All these variables were taken out of the dataset. This is important to identify as it could cause problems when fitting the model and interpreting the results.

**5c. Feature Selection**

Recursive Feature Elimination (RFE) was conducted to understand what are the most important variables to develop a predictive model. The variables that were left are the following:

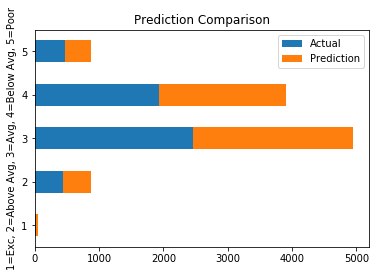
|  |  |  |  |
| --- | --- | --- | --- |
| **Model Key Variables** | | | |
| age | mentality aggression | movement balance | movement reactions |
| gk diving | physic | dribbling | mentality composure |
| height | preferred foot | player position | potential |
| international reputation | | | |

**5d. Model Selection**

For the exercise train and test data sets were created with a 70/30 split, and four potential classifiers were used for the exercise. K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Model accuracy was the metric selected to choose the appropriate model for the project.

Each model used a 10 fold for cross validation, and parameter tuning was used for at least one criteria to find the most accurate result for each one. At the end Random Forest was the classifier chosen, as it netted the highest accuracy at around 90%.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Accuracy Comparison** | | | |
| **KNN** | **RF** | **LR** | **SVM** |
| 0.8235 | 0.9003 | 0.8114 | 0.4586 |



**90% Accuracy**

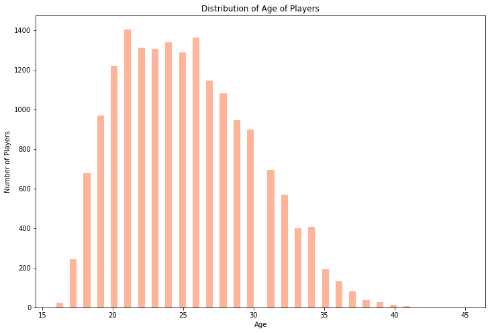
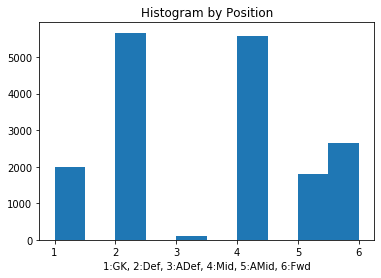
**5e. Prediction**

FIFA 20 dataset was uploaded and received the same data preprocessing, cleaning, missing data handling, and variable elimination. The Random Forest classifier was done to create the prediction, finally the results with some key data from the dataset were exported into a new one for clearer player selection.

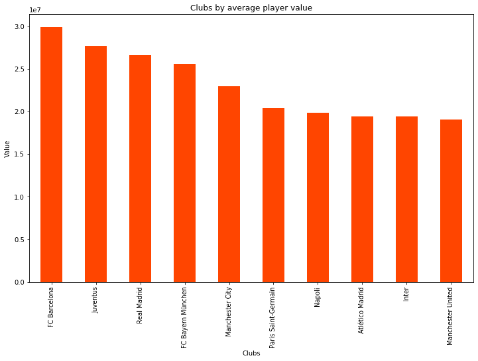
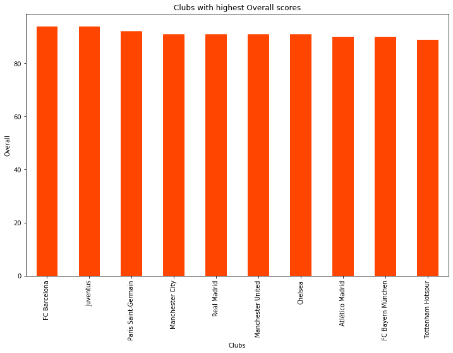
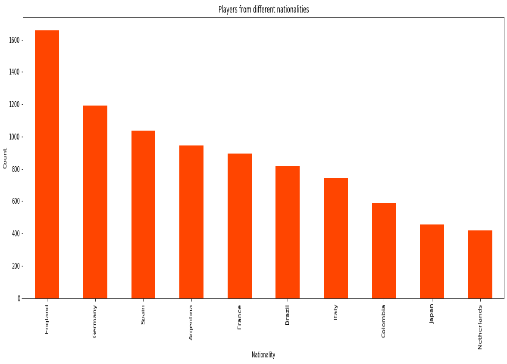
**6. Key Visualizations and Findings**

**Descriptive Analysis**

Analyzing the FIFA 19 dataset, most players in the pool are Defenders or Midfielders. They are also aged between 20 to 26 years old. As a result, finding different value players with similar skills is most likely in these positions.

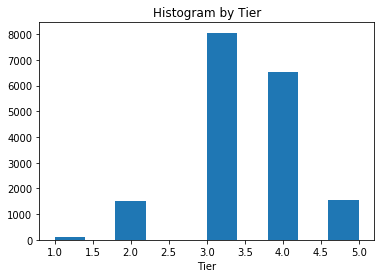
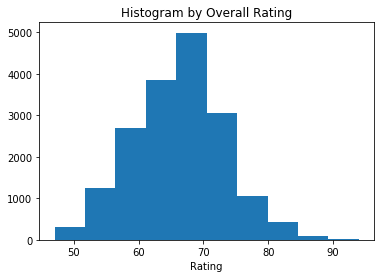


The top 3 nationalities come from England, Germany and Spain, and most of the best and highest value players play on the most powerful teams in Europe. Targeting solid players outside these nationalities and clubs could provide useful when finding similar skill players with lower value.

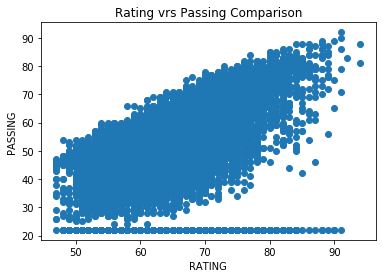
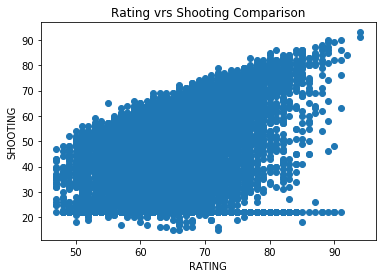
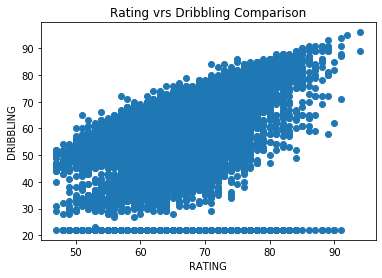


**Player Skills**

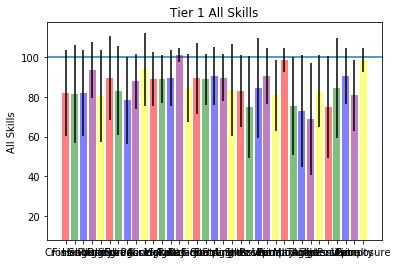
Players with an overall rating above 75 are scarce, but also vary along a big range of metrics. Replacing it with a Tier approach simplifies and broadens the scope of players to choose by grouping them into similar categories, therefore the importance of accurately predicting it.



Better overall player ratings are directed linked to broad skills such as shooting, passing, and dribbling.

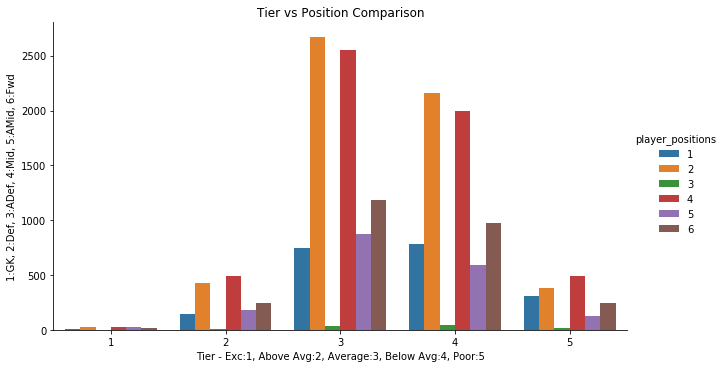
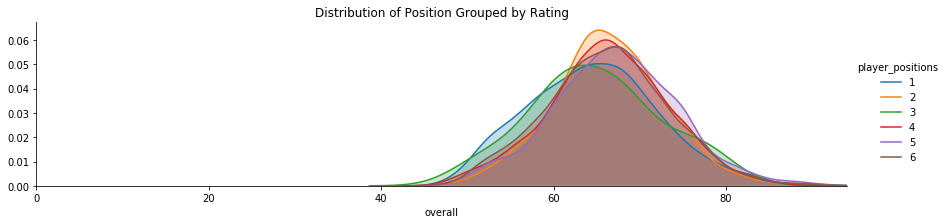
****

Specific skills that are strong on Tier 1 type players are short passing, control, stamina, and composure. Most of them have high power and skills metrics.

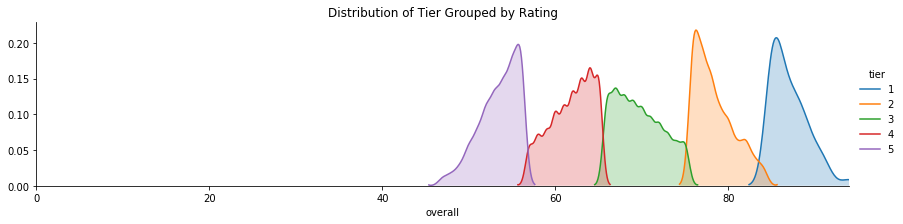


**Player Values**

There is a similar distribution of rated players by position. There is no need to jump on a specific position. Nevertheless, tier 1 players are evenly distributed among the Defender, Midfielder, and Attacking Midfielder positions. You should find more good players at this positions with different type of values.



Additionally, there are group of players that overlap tiers, creating an opportunity to find similar skill players at lower values. This could be a focus point when building the squad.



In average the highest player value is found in Tier 1, as it should occur. Nevertheless, there is a big standard deviation in the group, meaning that you could find similar good players at a fraction of a cost as the top ones.

|  |  |  |
| --- | --- | --- |
| **Tier** | **€ Value Mean** | **€ Value Standard Deviation** |
| 5 | 106,327.60 | 43,382.29 |
| 4 | 396,193.10 | 208,500.00 |
| 3 | 2,124,200.00 | 1,979,182.00 |
| 2 | 12,091,810.00 | 6,965,885.00 |
| 1 | 48,720,720.00 | 20,896,900.00 |

Overall finding is that there is value in defining player tiers. You are able to identify players with cheaper values that should perform similar to the most expensive and established football stars.

**7. Project Recommendation**

**Player Tiers**

As we analyze the data, it becomes apparent that being able to segment players in tiers becomes very useful. Specially in identifying potential opportunities, players valued lower but that should perform similar to more expensive stars. It provides valuable insight to support player selection and build a capable squad with the money allocated.

**Model Implementation**

In order to create the player tiers for the FIFA 20 data, the first step would be to run the model to quickly predict and group player performance. This would provide a tool to assess what is the right approach and actions to take when creating the best squad for the specific budget. The tool is flexible enough that it should work if budget constrains change. It has an estimated accuracy of around 90%, but personal strategy and preference are key to drive the most appropriate decision.

**Squad Selection**

There are different ways to select the players using the tool. Personal preference dictates it, as you may want to load on specific positions, or select a player despite value. Our recommendation would be to select the lowest value players from the top tiers in each position, avoiding going over the €275M threshold. That way you can make sure you have quality players without breaking the bank. Following this method the team we selected, with a total budget of ***€268M***, is the following:

**Recommended FIFA 20 Squad**