Identifying Top Players and Significant Player Attributes Based on Player Position for European Football

Springboard Intermediate Data Science: Python

Capstone Project Final Report

Prepared by: Ruhama Ahale (ruhama.ahale@gmail.com)

Mentor: Raghunandan Patthar

October 2017

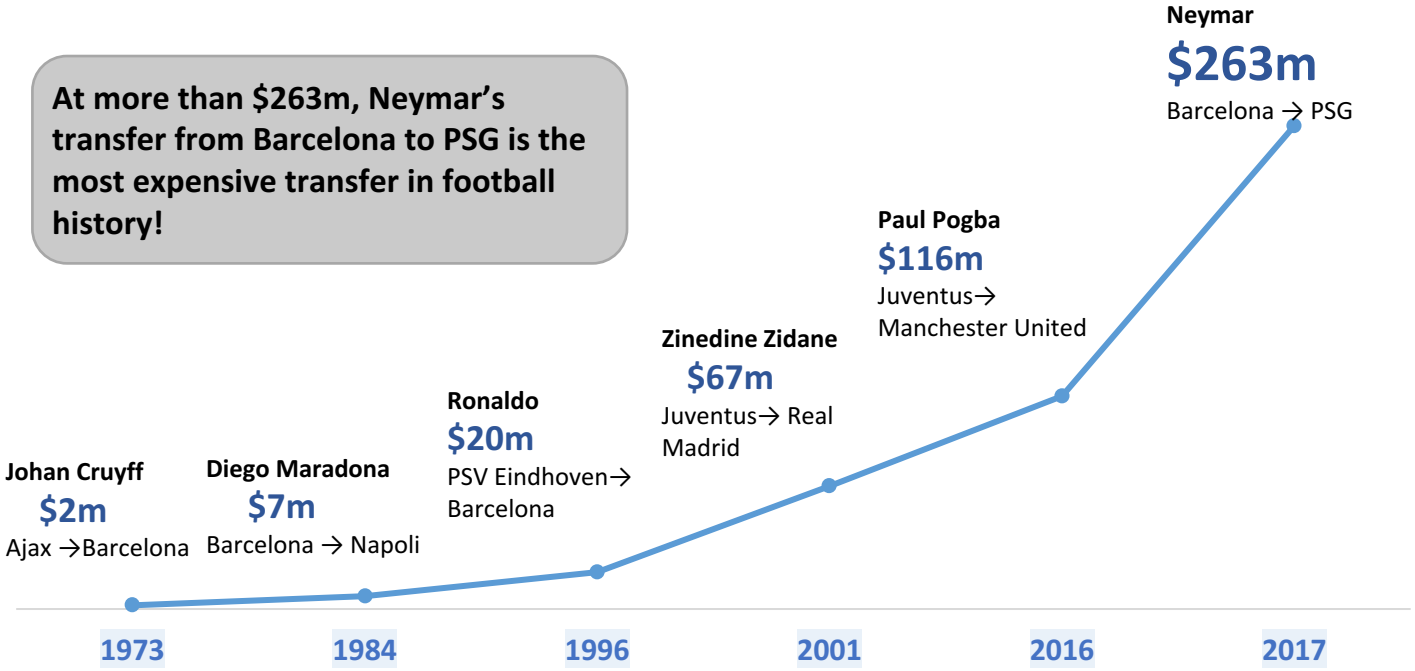
TABLE OF CONTENTS

INTRODUCTION

PROBLEM

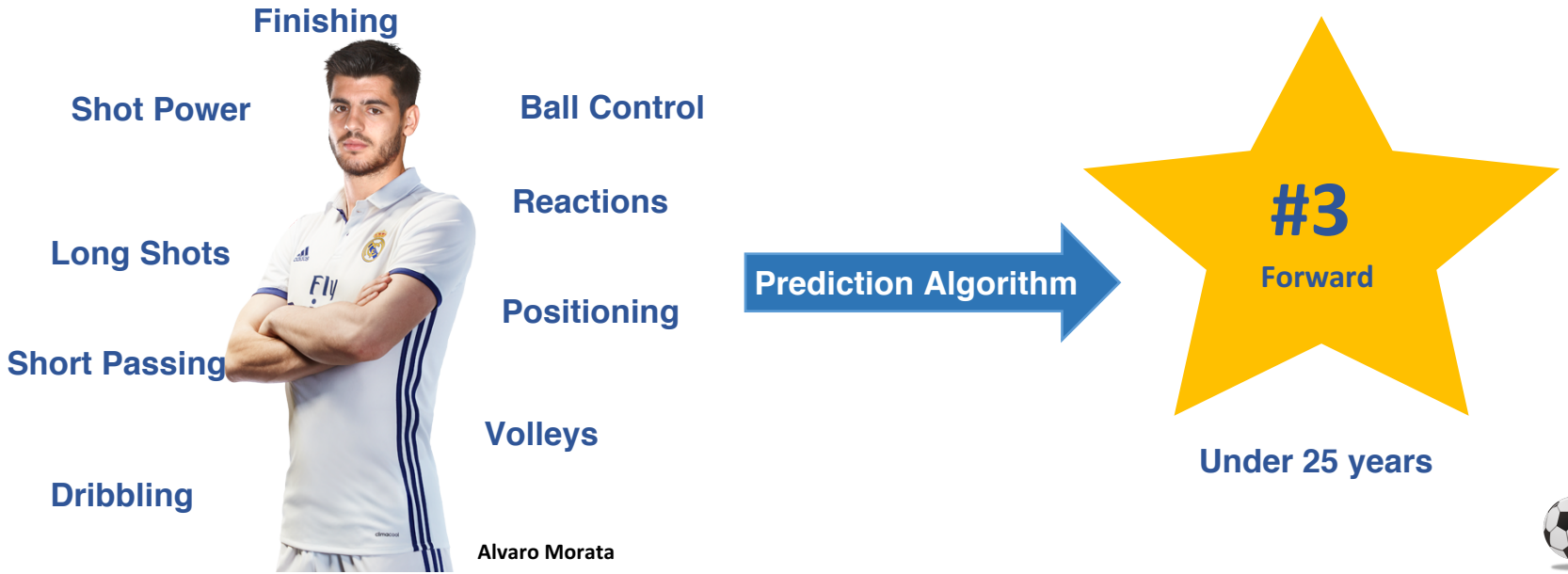
Player transfers in association football are multimillion dollar deals, every year clubs pay millions of dollars to sign new players. The client, a European football club wants a list of top players for each playing position so that they can shortlist desirable candidates from the players available on the transfer list.

At more than $263m, Neymar's transfer from Barcelona to PSG is the most expensive transfer in football history!



REQUIREMENT

- The client wants a model that can predict the aggregated overall rating of a player at each age of the player. In addition to the **list of top players at playing positions** they also want to know **the significant attributes that affect the player rating for each position**
- Player ratings are calculated by FIFA based on player attribute stats as well as other factors such as the league the player is associated with, the specific circumstances which resulted in an attribute score and various other factors
- Our aim is to **predict these ratings based only on the player attribute stats available to us from the FIFA video game data**



DATA SOURCE

- The data we will use for this problem is sourced from several websites such as
  Football-Data-MX-API: scores, lineup, team formation and events
  Sofifa: players and teams attributes from EA Sports FIFA games, FIFA series and all FIFA assets property of EA Sports;
- It is curated by Hugomathien and made available on Kaggle [here]
- Data available from 2008 to 2016 in form of SQLite database

TABLES:

- **Match**: Match details data for each match like match_date, home_team, opponent_team, player_ids of players playing in the match, player_id at each playing coordinate
- **Player**: Player details like player_id, player_name, birthdate, height, weight
- **Player_Attributes**: Attributes sourced from several sources for each match player has played like player_id, preferred_foot, attacking_work_rate, overall_rating, crossing, finishing, heading_accuracy, short_passing etc.
- **Country:** country_id, country_name
- **League:** id, country_id, league_name

METHODOLOGY

DATA WRANGLING

- Clean
- Merge
- Handle Missing Values
- Aggregate

DATA EXPLORATION

- Data Distribution
- Hypothesis Testing: Difference Between Groups, Normality, Linearity, Multicollinearity

DATA MODELING

- Feature Selection
- Feature Reduction (PCA)
- Regression Models: Linear, Polynomial, Ridge, Lasso, SVR
- Model Comparison
- Prediction

DATA WRANGLING

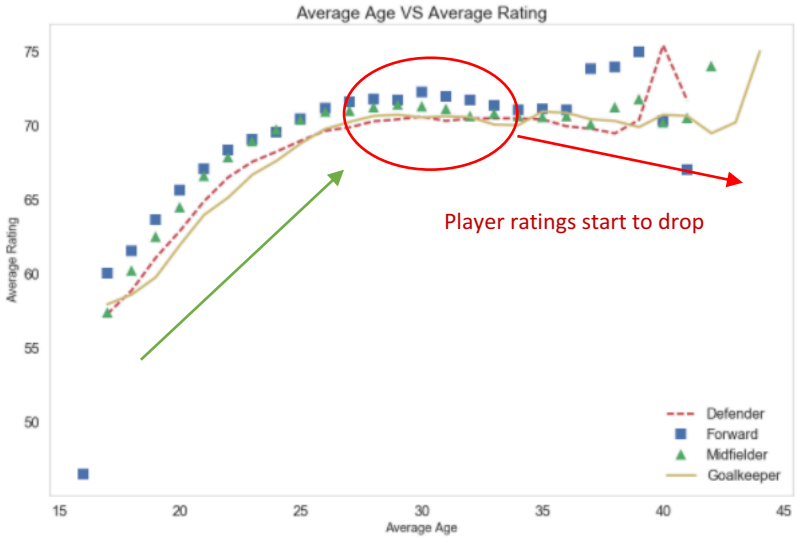
- Player positions are calculated for each player using Match data
- Player, Player Attributes, Positions files are merged to get all player data in a single data table, the variables in this table can be seen in the adjoining image
- Less than 5% of the data was missing i.e. some features were not available for some players, also data from the season 2008-09 was very inconsistent. These missing values were removed from the final dataset
- Data was aggregated using the mean function for each player at each age of the player because the model requirement is to get predicted scores for a player at players current age

```
Data columns (total 39 columns):
player_api_id         52725 non-null int64
player_name           52725 non-null object
player_position       52725 non-null object
overall_rating        52725 non-null float64
age                   52725 non-null float64
height                52725 non-null float64
weight                52725 non-null int64
finishing             52725 non-null float64
heading_accuracy      52725 non-null float64
short_passing         52725 non-null float64
volleys               52725 non-null float64
dribbling             52725 non-null float64
curve                 52725 non-null float64
free_kick_accuracy    52725 non-null float64
long_passing          52725 non-null float64
ball_control          52725 non-null float64
acceleration          52725 non-null float64
sprint_speed          52725 non-null float64
agility               52725 non-null float64
reactions             52725 non-null float64
balance               52725 non-null float64
shot_power            52725 non-null float64
jumping               52725 non-null float64
stamina               52725 non-null float64
strength              52725 non-null float64
long_shots            52725 non-null float64
aggression            52725 non-null float64
interceptions         52725 non-null float64
positioning           52725 non-null float64
vision                52725 non-null float64
penalties             52725 non-null float64
marking               52725 non-null float64
standing_tackle       52725 non-null float64
sliding_tackle        52725 non-null float64
gk_diving             52725 non-null float64
gk_handling           52725 non-null float64
gk_kicking            52725 non-null float64
gk_positioning        52725 non-null float64
gk_reflexes           52725 non-null float64
dtypes: float64(35), int64(2), object(2)
```

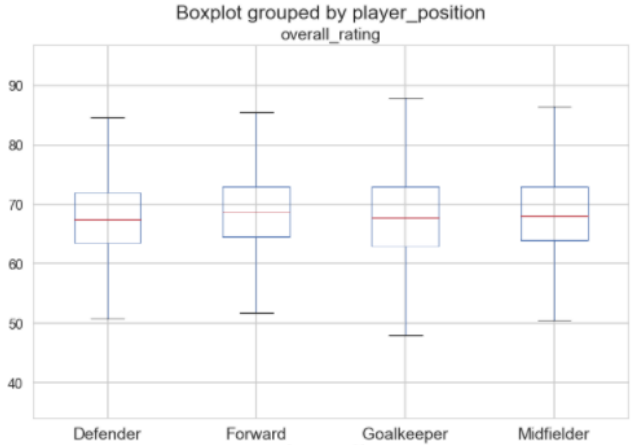
Player data information

DATA EXPLORATION

- It is assumed that a players playing position and players age affects the players rating.
- We test these assumptions and find that player rating tends to increase until the player reaches approximately 30, while after 30 the growth stops and eventually ratings go down



- We can use the entire dataset available to build a model to predict player rating, however, our assumption that player position affects players rating is correct. We tested the hypothesis that there is no significant difference between mean ratings of different position groups using the one way ANOVA test and based on results of the test we rejected the hypotheses
- As a result, we will build prediction model separately for each position



Boxplot shows mean difference between groups

DATA MODELING

FEATURE SELECTION

- Players at different positions have different skill sets, based on the available attribute set we identify features that are significantly affecting the player rating
- As the dataset is very large with many features (36). We select features for each player position such that if there is no significant relationship between player rating and feature we discard the feature.
- The selected features are scaled to increase modeling accuracy
- Principal Component Analysis is performed on the scaled feature data to reduce dimensionality and the new components are selected basis scree plot.
- These new components are used in two of the models: Linear Regression and Polynomial Regression as these models do not work well with high dimensions

Defender:

```
Insignificant features are:
```

	features	r_score	r2_score	pvalue
34	gk_reflexes	0.009236	0.001349	0.2215
33	gk_positioning	0.003310	0.000012	0.6613
31	gk_handling	-0.010202	0.000536	0.1769

Forward:

```
Insignificant features are:
```

	features	r_score	r2_score	pvalue
31	gk_handling	0.023150	0.000536	0.0195
30	gk_diving	0.021562	0.000465	0.0296
33	gk_positioning	0.003423	0.000012	0.7299

Midfielder:

```
Insignificant features are:
```

	features	r_score	r2_score	pvalue
33	gk_positioning	0.003289	0.000012	0.6406
34	gk_reflexes	0.002207	0.001349	0.754
31	gk_handling	0.001654	0.000536	0.8144
1	height	-0.016278	0.000819	0.0208

Goalkeeper:

```
Insignificant features are:
```

	features	r_score	r2_score	pvalue
3	finishing	-0.016550	0.583881	0.2484
6	volleys	-0.019294	0.583784	0.1784
21	long_shots	-0.020163	0.532845	0.1596
29	sliding_tackle	-0.028369	0.007979	0.0478
7	dribbling	-0.028393	0.526532	0.0477

The respective insignificant features for each of these positions are excluded from the data modeling for the position

MODEL COMPARISON

- All datasets are split into train and test data (80-20 split)
- Models are optimized on train data by obtaining best parameters and performing K-fold Cross Validation
- Best fitting model is selected for each position by comparing the test accuracy and root mean square error of the model
- The following modeling techniques were used:
  - **Linear Regression**
  - **Polynomial Regression**
  - **Ridge Regression**
  - **Lasso Regression**
  - **Support Vector Regression (SVR)**
- For each position Polynomial Regression had overfitting issues and was thus not considered
- SVR was the best fit model for predicting ratings for each player position. As such, the predicted player ratings for last ages (in data) of players were computed using SVR.

Defender Model Comparison

	Model	Test_Accuracy
0	Linear_Regression	75.32
1	Ridge_Regression	89.08
2	Lasso_Regression	89.06
3	SVR	97.79

Forward Model Comparison

	Model	Test_Accuracy
0	Linear_Regression	56.22
1	Ridge_Regression	91.56
2	Lasso_Regression	91.64
3	SVR	98.21

Midfielder Model Comparison

	Model	Test_Accuracy
0	Linear_Regression	65.71
1	Ridge_Regression	86.60
2	Lasso_Regression	86.61
3	SVR	97.06

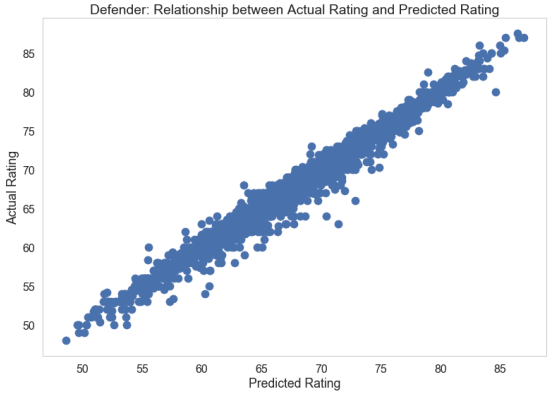
Goalkeeper Model Comparison

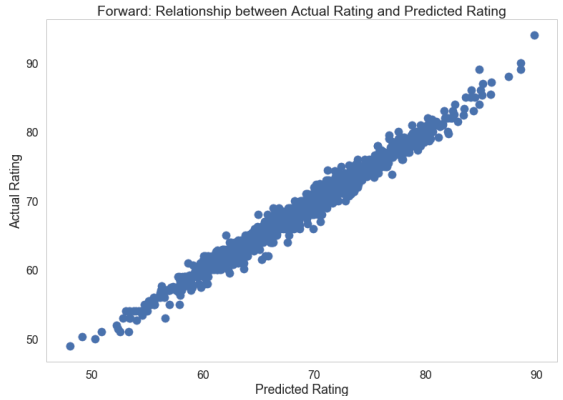
	Model	Test_Accuracy
0	Linear_Regression	-266.57
1	Ridge_Regression	87.14
2	Lasso_Regression	86.61
3	SVR	97.06

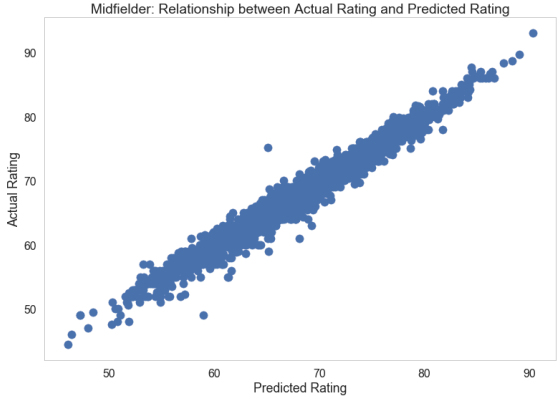
Support Vector Regression (SVR) is the best fitting model

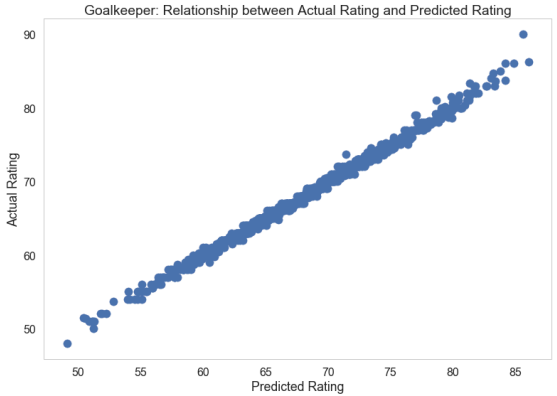
PREDICTION

The scatter plots of the actual vs predicted ratings for all four positions show almost negligible dispersion indicating that our predictions have high accuracy









RESULTS

- We identified the top attributes which have very high correlation with the ratings of players for each position
- We identified top players by position overall and also top players by position under age of 25 years
- The list of these top players is available in the appendix (slide x)
- Our models have been successful in predicting player rating using only the attribute stats available with more than 97% accuracy
- Our predicted ratings are for 2016*, some of our predicted top players such as Neymar, Lukaku, Morata were transferred this season. This shows how useful our list can be for player selection
- The final list can be filtered by age, country, league or other player details based on clients requirement. This will help narrow down the search as desired

MOST SIGNIFICANT FEATURES

DEFENDER

- Standing Tackle, Interceptions, Marking, Sliding Tackle and Reaction

FORWARD

- Ball Control, Reactions, Positioning, Finishing, Volleys, Shot Power, Long Shots, Short Passing and Dribbling

MIDFIELDER

- Ball Control, Reactions, Vision, and Short Passing

GOALKEEPER

- Goalkeeping positioning, Goalkeeping Reflexes, Goalkeeping Diving and Goalkeeping Handling

*2017 data will be added to the analysis in the future

LIMITATIONS

There are certain limitations in the data that reduce the robustness of the models we have developed:

- Not all historical data is available for each player. Knowing that the FIFA Player Ranking is based on historical data of player, we do not have this data for each player. A lot of information will not be learned by the model due to the absence of this data.
- Data related to ranks of leagues, teams and countries is also not available in this data. This information would have made the model more accurate as these factors play an important role in determining player ranking.
- There is not enough historical data available for each player, if it were, a time series model could be built to predict the future ratings of players
- There was country and league data missing for few players and more than 20% players from our players' data had not been active in the last two years. These players were excluded from featuring in the top players lists.

RECOMMENDATIONS

- The player rating increases with age and for Defender, Midfielder and Forwards around the age range of 30-35 the rating starts to drop
- For goalkeepers, the trend is different with ratings remaining high from ages 25 onwards
- The general assumption that player rating falls as age increases is true only after age range of 30-33 years. This information can be used on the best players given by the models to make better selections
- We have identified the current best players and this list can be used to filter out potential new additions to the club
- The top significant features that are identified in this study should be further analyzed and these skills should be identified early in newer players to improve their performance

FUTURE WORK

Add FIFA 2017 data to get current player list

Analyze effect of age, nationality on player attributes

Identify patterns in attributes of top players

Make predictions for more specific player positions

Analyze historical performance of players to make future predictions

Use attribute patterns of top players to identify potentially high performing newcomers

Thank You 

APPENDIX

TOP 10 DEFENDERS

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
80562	Defender	Thiago Silva	86	87	2	France Ligue 1	France	31
30962	Defender	Sergio Ramos	86	87	2	Spain LIGA BBVA	Spain	30
30894	Defender	Philipp Lahm	86	87	2	Germany 1. Bundesliga	Germany	32
36183	Defender	Jerome Boateng	85	87	4	Germany 1. Bundesliga	Germany	27
56678	Defender	Diego Godin	85	85	4	Spain LIGA BBVA	Spain	30
36388	Defender	Mats Hummels	84	86	8	Germany 1. Bundesliga	Germany	27
121633	Defender	David Alaba	84	85	8	Germany 1. Bundesliga	Germany	24
31306	Defender	Branislav Ivanovic	84	80	8	England Premier League	England	32
19327	Defender	Miranda	84	85	8	Italy Serie A	Italy	32
39027	Defender	Vincent Kompany	84	85	8	England Premier League	England	30

TOP 10 DEFENDERS
UNDER 25

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
121633	Defender	David Alaba	84	85	1	Germany 1. Bundesliga	Germany	24
115591	Defender	Ricardo Rodriguez	82	82	2	Germany 1. Bundesliga	Germany	24
282674	Defender	Daniel Carvajal	81	81	3	Spain LIGA BBVA	Spain	24
230982	Defender	Raphael Varane	81	82	3	Spain LIGA BBVA	Spain	23
267365	Defender	Marquinhos	80	80	6	France Ligue 1	France	21
184999	Defender	Shkodran Mustafi	80	82	6	Spain LIGA BBVA	Spain	24
188555	Defender	Stefan de Vrij	80	81	6	Italy Serie A	Italy	24
411617	Defender	Aymeric Laporte	80	81	6	Spain LIGA BBVA	Spain	21
213485	Defender	Matija Nastasic	79	79	9	Germany 1. Bundesliga	Germany	23
195299	Defender	Serge Aurier	79	80	9	France Ligue 1	France	23

TOP 10 FORWARDS

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
30981	Forward	Lionel Messi	89	94	1	Spain LIGA BBVA	Spain	29
35724	Forward	Zlatan Ibrahimovic	88	89	2	France Ligue 1	France	34
40636	Forward	Luis Suarez	88	90	2	Spain LIGA BBVA	Spain	29
37412	Forward	Sergio Aguero	87	88	4	England Premier League	England	28
38817	Forward	Carlos Tevez	85	85	6	Italy Serie A	Italy	32
164684	Forward	James Rodriguez	85	87	6	Spain LIGA BBVA	Spain	25
93447	Forward	Robert Lewandowski	85	87	6	Germany 1. Bundesliga	Germany	27
30829	Forward	Wayne Rooney	85	85	6	England Premier League	England	30
19533	Forward	Neymar	84	89	9	Spain LIGA BBVA	Spain	24
26166	Forward	Karim Benzema	84	86	9	Spain LIGA BBVA	Spain	28

TOP 10 FORWARDS
UNDER 25

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
19533	Forward	Neymar	84	89	1	Spain LIGA BBVA	Spain	24
325916	Forward	Paulo Dybala	82	79	2	Italy Serie A	Italy	22
181276	Forward	Romelu Lukaku	80	81	5	England Premier League	England	23
213501	Forward	Alvaro Morata	80	80	5	Italy Serie A	Italy	24
303824	Forward	Memphis Depay	80	80	5	England Premier League	England	22
241825	Forward	Francisco Alcacer	80	80	5	Spain LIGA BBVA	Spain	23
364520	Forward	Domenico Berardi	80	81	5	Italy Serie A	Italy	22
194165	Forward	Harry Kane	80	81	5	England Premier League	England	23
354467	Forward	Yannick Ferreira-Carrasco	79	79	9	Spain LIGA BBVA	Spain	23
202443	Forward	Vincent Aboubakar	79	78	9	Portugal Liga ZON Sagres	Portugal	24

TOP 10 MIDFIELDERS

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
30893	Midfielder	Cristiano Ronaldo	90	93	1	Spain LIGA BBVA	Spain	31
30834	Midfielder	Arjen Robben	89	89	2	Germany 1. Bundesliga	Germany	32
107417	Midfielder	Eden Hazard	88	88	3	England Premier League	England	25
30955	Midfielder	Andres Iniesta	87	88	4	Spain LIGA BBVA	Spain	31
37459	Midfielder	David Silva	86	86	6	England Premier League	England	30
30924	Midfielder	Franck Ribery	86	86	6	Germany 1. Bundesliga	Germany	33
31921	Midfielder	Gareth Bale	86	87	6	Spain LIGA BBVA	Spain	27
129944	Midfielder	Marco Reus	86	86	6	Germany 1. Bundesliga	Germany	27
39854	Midfielder	Xavi Hernandez	85	86	9	Spain LIGA BBVA	Spain	35
154257	Midfielder	Sergio Busquets	85	86	9	Spain LIGA BBVA	Spain	28

TOP 10 MIDFIELDERS
UNDER 25

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
248453	Midfielder	Paul Pogba	84	86	1	Italy Serie A	Italy	23
184536	Midfielder	Philippe Coutinho	83	84	2	England Premier League	England	24
177714	Midfielder	Mario Goetze	83	84	2	Germany 1. Bundesliga	Germany	24
184533	Midfielder	Koke	82	83	6	Spain LIGA BBVA	Spain	24
191315	Midfielder	Isco	82	84	6	Spain LIGA BBVA	Spain	24
243164	Midfielder	Julian Draxler	82	82	6	Germany 1. Bundesliga	Germany	23
190972	Midfielder	Marco Verratti	82	84	6	France Ligue 1	France	23
157723	Midfielder	Christian Eriksen	82	83	6	England Premier League	England	24
174850	Midfielder	Erik Lamela	81	79	9	England Premier League	England	24
242709	Midfielder	Roberto Firmino	81	81	9	England Premier League	England	24

TOP 10 GOALKEEPERS

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
182917	Goalkeeper	David De Gea	86	86	1	England Premier League	England	25
27299	Goalkeeper	Manuel Neuer	85	90	2	Germany 1. Bundesliga	Germany	30
170323	Goalkeeper	Thibaut Courtois	84	86	4	England Premier League	England	23
42422	Goalkeeper	Samir Handanovic	84	83	4	Italy Serie A	Italy	32
30859	Goalkeeper	Petr Cech	84	86	4	England Premier League	England	34
215168	Goalkeeper	Bernd Leno	83	83	7	Germany 1. Bundesliga	Germany	24
26295	Goalkeeper	Hugo Lloris	83	85	7	England Premier League	England	29
31432	Goalkeeper	Joe Hart	83	84	7	England Premier League	England	29
30717	Goalkeeper	Gianluigi Buffon	83	84	7	Italy Serie A	Italy	38
37421	Goalkeeper	Claudio Bravo	82	83	10	Spain LIGA BBVA	Spain	32

TOP 10 GOALKEEPERS
UNDER 25

player_api_id	player_position	player_name	predicted_overall_rating	overall_rating	player_rank	league_name	country_name	age
170323	Goalkeeper	Thibaut Courtois	84	86	1	England Premier League	England	23
215168	Goalkeeper	Bernd Leno	83	83	2	Germany 1. Bundesliga	Germany	24
184554	Goalkeeper	Marc-Andre ter Stegen	81	82	3	Spain LIGA BBVA	Spain	24
181910	Goalkeeper	Mattia Perin	80	81	5	Italy Serie A	Italy	23
177126	Goalkeeper	Jan Oblak	80	81	5	Spain LIGA BBVA	Spain	23
212815	Goalkeeper	Timo Horn	80	80	5	Germany 1. Bundesliga	Germany	23
245555	Goalkeeper	Geronimo Rulli	79	79	7	Spain LIGA BBVA	Spain	24
288880	Goalkeeper	Jack Butland	79	78	7	England Premier League	England	23
287894	Goalkeeper	Loris Karius	78	79	9	Germany 1. Bundesliga	Germany	23
210164	Goalkeeper	Alphonse Areola	77	77	10	Spain LIGA BBVA	Spain	23