**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**

1. INTRODUCTION
2. DATA ACQUISITION AND CLEANING
   1. DATA SOURCE
   2. HANDLING MISSING VALUES
3. METHODOLOGY
   1. DATA PREPARATION
   2. DATA EXPLORATION
   3. MODELING TECHNIQUES
4. POSITIONAL DATA INSIGHTS
   1. FEATURE SELECTION
   2. DATA PREPARATION
   3. FEATURE REDUCTION
   4. DATA MODELING
   5. MODEL COMPARISON
   6. PREDICTION
5. RESULTS
6. LIMITATIONS
7. FUTURE WORK
8. CLIENT RECOMMENDATIONS
9. INTRODUCTION

The client, a European Football Club wants a model to predict the aggregated overall rating of a player at each age of the player. They want to identify the current top players for each playing position and the significant attributes that affect the player rating for each position. They will use this list of best players to consider during transfer season. In case they need to make a transfer, they will bring in a replacement, a major part of player selection depends upon which position the player will play at. Depending on the need of the team, the managers will consider the up and coming players who will suit their need.

Based on an interview with Mueller-Moehring who is responsible for rating players in FIFA games every year [[here](https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings/)], the 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 goal is to fit a model that predicts the overall rating using only the player attributes available to us in the data source.

1. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCE:

The data we will use for this problem is sourced from several websites such as

http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events

http://sofifa.com/ : 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](https://www.kaggle.com/hugomathien/soccer)]

We want to identify top players playing at each of the four positions - Forward, Defender, Midfielder and Goalkeeper from the dataset of players playing for different clubs in European Football since 2008 up to 2016.

Below are details of each table:

1. Match: Match data with each row having - match\_date, home\_team, opponent\_team, player\_ids of players playing in the match, player\_id at each playing coordinate

2. Player: Player details- player\_id, player\_name, birthdate, height, weight

3. Player\_Attributes: Attributes sourced from several sources for each match player has played - player\_id, preferred\_foot, attacking\_work\_rate, overall\_rating, crossing, finishing, heading\_accuracy, short\_passing etc.

4. Country: country\_id, country\_name

5. League: id, country\_id, league\_name

Data is collected from this database and processed to merge some of these files to get player positions data and player league and country data in a more accessible format. It is not necessary that a player plays at the same position for each match, so we choose the players favored position using the mode function to calculate the player position. These new tables are then pushed back into the database and this new SQLite database is used as source for this analysis. [[Here](https://github.com/ruhama-ahale/soccer_project/blob/master/Capstone_Final_Report/Data_Wrangling_Capstone_Project.ipynb)] is the link to the ipynb file which contains the data manipulations code in detail.

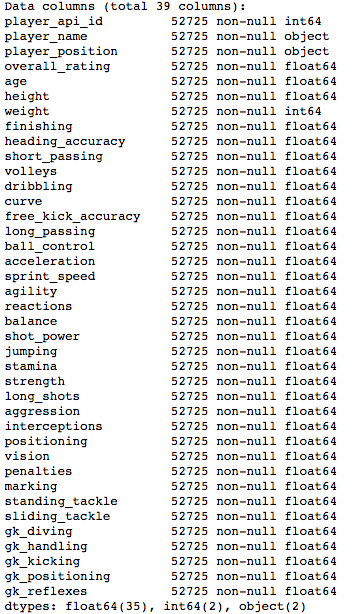


Fig 1.1.1

Figure 1.1 shows the information about the data in the consolidated dataset. We use this data for data exploration and analysis.

* 1. HANDLING MISSING VALUES

Since data from 2008-09 season is inconsistent, we exclude that from our analysis. From the remaining data, the proportion of missing values is very small 2.6%, we ignore the missing values and delete them from the dataset as they are less than 5%. The data analysis and modeling is done on this data.

However, in the results, we need to merge the prediction data with the table having players’ country and league detail. This table has more missing data, mainly any player who hasn’t played since Jan 2013 is not available in this table. It makes sense to show only active players in the list of top players so we exclude these missing values from the final output.

1. METHODOLOGY

This is a prediction problem where using the independent data i.e. player attributes we must predict the dependent variable i.e. overall rating. Fig 3.1 depicts a flow chart of the methodology used for this study.

Fig 3.1

* 1. DATA PREPARATION

We have already merged the files and removed missing values using the raw data. To get predicted ratings for players at the next age of player, we want data for ratings and attributes of a player at a given age. The data that we have has ratings and attributes at multiple matches for multiple ages of a player. So, for one age of a player e.g. 25, there could be more than one match details. To get one data point for each age of a player we aggregate the features at that age using the average function.

Data Description:

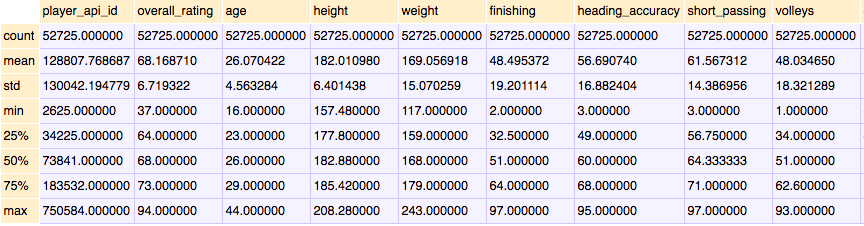


Fig 3.1.1

* 1. DATA EXPLORATION

**Player Age Distribution:**

Since the relationship of the player age and player rating is of interest to us, as there is a general assumption in football that player rating decreases with age of player, let us explore if this assumption is true for each of the playing positions

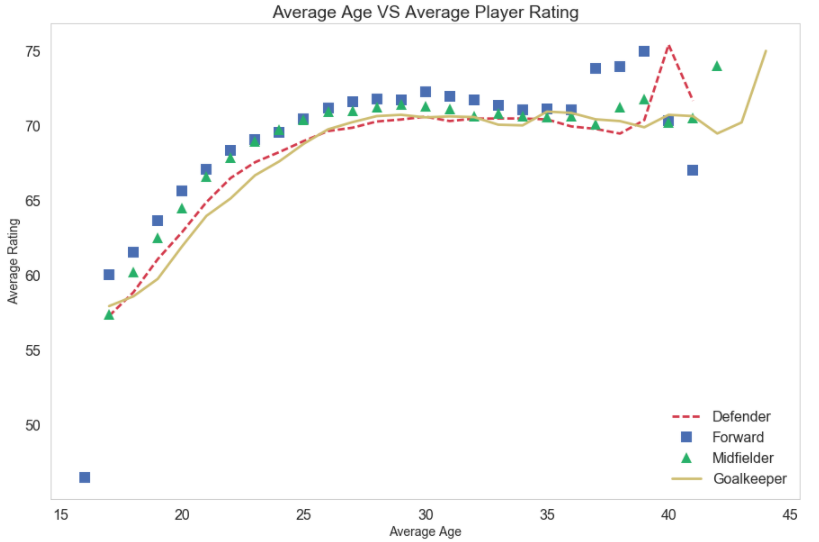


Fig 3.2.1

From the above plot, we can see that the player rating increases with age and for Defender, Midfielder and Forwards around the age range of 30-35 the rating starts to drop. While for goalkeepers, the trend is different with ratings remaining high from ages 25 onwards. So, the assumption that player rating falls as age increases is true only after age range of 30-33 years.

**Plot the distribution of ages for players by positions**

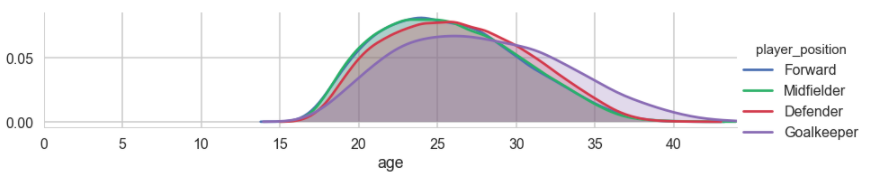


Fig 3.2.2

From the above plot, we can see that the distribution of player age is not normally distributed, with the most players for Defender, Midfielder and Forward positions being around the age range of 22-25. While for goalkeepers, most players are in the age range of 23-32 and there are more older goalkeepers than other players

**Difference in ratings between positions**

To predict the player rating we can separate the data into train and test datasets and fit a regression model to the data to find a model with a good fit. However, we know that the data has 36 attributes for each player and according to our assumptions about the game, we think that there is a difference between ratings based on which position the player plays at. If this is true, it would make sense for us to separate the data by positions and fit separate models on each position data. For this, we test the following hypothesis:

H0: There is no significant difference between the mean ratings of the different position groups

VS

H1: There is significant difference between the mean ratings of the different position groups

We test the difference between the means using one way anova and at 5% l.o.s we reject the null hypotheses for the alternate. Check Fig 3.2.3, p-value = 6.67….e-54

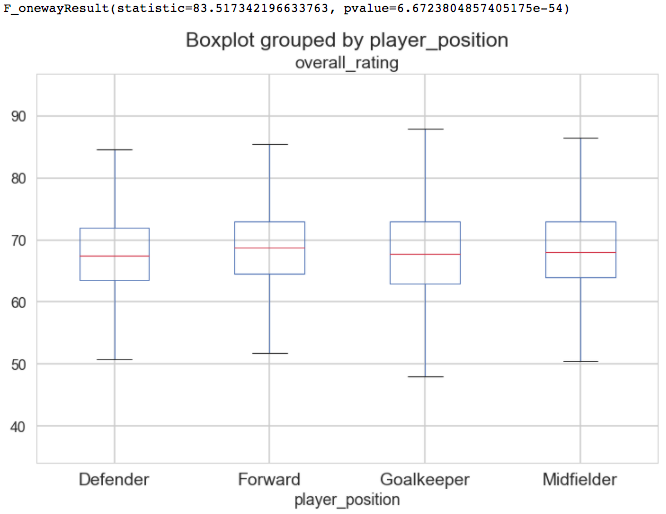


Fig 3.2.3

We carry out Tukey's test using the pairwise\_tukeyhsd() function in the statsmodels.stats.multicomp library:

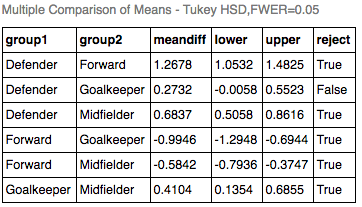


Fig. 3.2.4

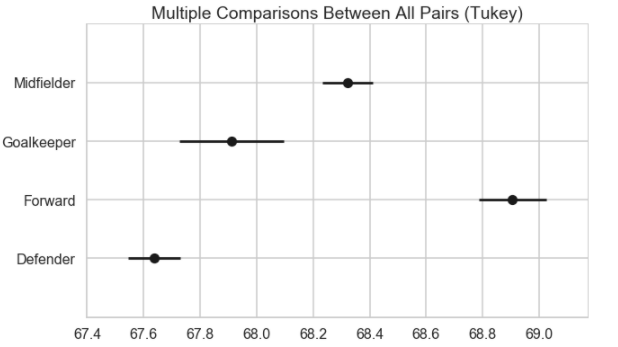


Fig 3.2.5

The output of the Tukey test shows the average difference, a confidence interval as well as whether you should reject the null hypothesis for each pair of groups at the given significance level. In this case, the test suggests we reject the null hypothesis for 5 pairs, with an exception of the groups Defender-Goalkeeper. The 95% confidence interval plot reinforces the results visually: only Defender and Goalkeeper groups' confidence intervals overlap.

We have enough evidence to conclude that there is difference in ratings between the positional groups and we go forward with our suggestion to fit separate models for each position.

**Testing Normality, Linear Relationship and Multicollinearity:** We tested the data for checking distribution of available features and using the normaltest() function from the scipy.stats.mstats library we concluded that most of the variables fail the normality test. Overall Rating, Height, Weight, Marking and Sliding Tackle are the only variables which are normally distributed. Hence, to fit regression models to this data, we will need to normalize these features.

To check for linear relationship between each of the attributes and the dependent variable ‘overall\_rating’ we made scatter plots. According to these plots the relationship between most of the features and the rating is not linear but curvilinear, according to this a better fit would be Ridge Regression or Support Vector Regression (SVR)

To test for multicollinearity between the variables, we made a heat map of correlation between the variables, the result was as shown in Fig 3.2.6. As we can see most these variables are correlated, so there is multicollinearity present in the data, to deal with this we will use PCA to find independent components which can be used to predict overall rating. Note that the goalkeeping attributes have high negative correlation with the other attributes. We need to select the features with significance for each of the positions to build an efficient model

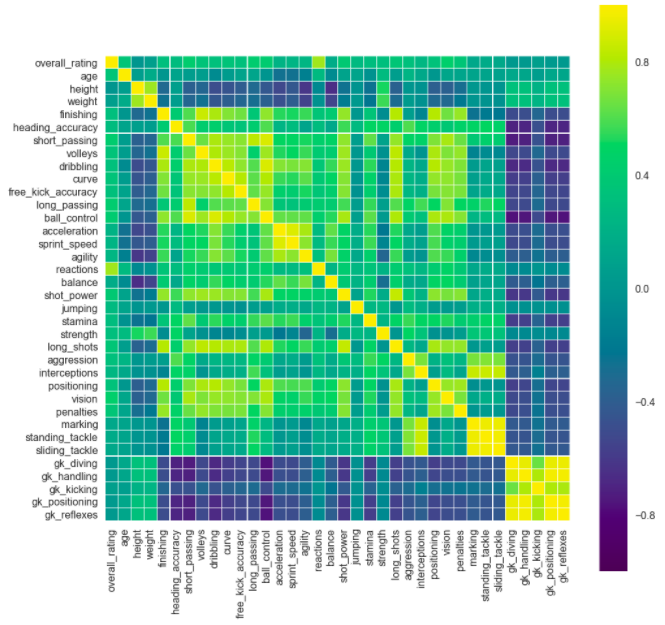


Fig. 3.2.6

* 1. MODELING TECHNIQUES

**Feature Selection and Dimensionality Reduction:** The dataset is very large with many features (36). Feature selection was done basis player position where if there is no significant relationship between player rating and feature we discard the feature. For e.g. For non-goalkeeper positions some goalkeeper

related features aren't significant so we discard them. 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 linear and polynomial regression models. While for Ridge, Lasso and SVR we used the original feature data after scaling as these methods inherently deal with multidimensionality issues

**Model Selection**: We split the data into training (~80%) and test (~20%) and then find the best fitting model for this transformed (train) data to predict the overall ratings for the next ages of the players (test data). We fit the following models and find one with the best fit to predict the rating at the next age for each player:

**Linear Regression**: Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables

**Polynomial Regression**: Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x

**Ridge Regression**: Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable

**Lasso Regression**: The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the regression coefficients

**Support Vector Regression**: Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. Firstly, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem

We compare the models and select the best fit to predict player ratings. Using these predicted ratings, we find the top x (client can change x as required) players for each position with the highest rating. This will be the data reported to the client to select a suitable player for the upcoming transfer season. As per the requirement, we can filter the players by age, country, league or other performance factors to find the best potential players the club will be interested in.

1. POSITIONAL DATA INSIGHTS

4.1. FEATURE SELECTION

For feature selection for the Position datasets, we find the r, r-squared, p-value for measuring the relationship of each independent feature with the dependent variable 'overall\_rating'. The Null Hypothesis is

H0: There will be no significant prediction of overall rating by feature VS H1: There will be significant prediction of overall rating by feature

We test this hypothesis at l.o.s. 1% and use only the significant features for model building

Null Hypotheses is **true** for the following features for the corresponding position data and they are excluded as insignificant features:



Fig 4.1

* 1. DATA PREPARATION

We separate the data into train and test datasets such that 80% data is training and 20% is testing. However, as our aim is to predict rating for current age of player we take maximum age of each player and separate it as test data. For each of the positions the proportion of this test data is found to be approximately 20%

The train-test datasets are then split into x and y datasets such that x is a subset of all independent features to be used for modeling and y is the target variable- player rating. The x datasets are then scaled using StandardScaler() function from the preprocessing library.

* 1. FEATURE REDUCTION

We use Principal Component Analysis (PCA) method for feature reduction. PCA is a learning method where the input data is to principal components to lower the dimensions of the data. A good rule of thumb with PCA is that we should be able to explain 95% variance with the reduced dimensions. If we plot number of components vs variance retained we should see that the explained variance increases steadily and then saturates, see Fig 4.3.1 for the scree plot of Defender dataset:

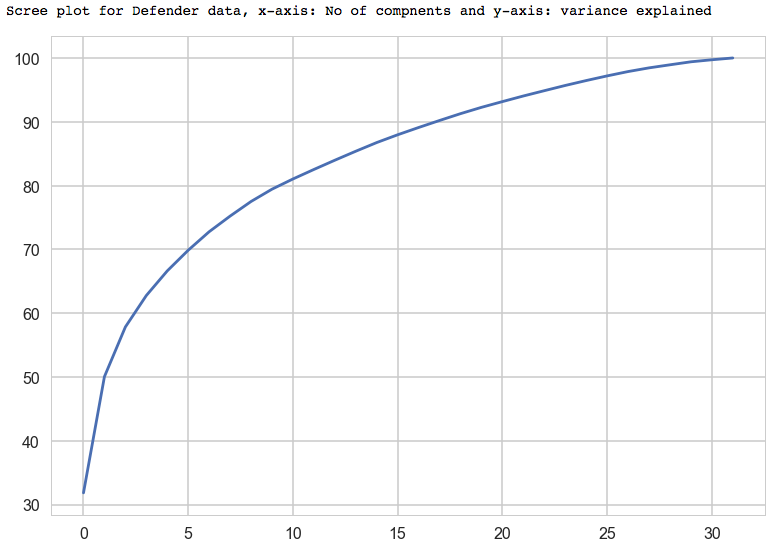


Fig 4.3.1

As we can see, % of variance in data is explained by 24 components, we select these 24 components for model building wherever feature reduction is required. Now we have 24 variables instead of 32.

Similarly, we use PCA to find reduced components for Forward, Midfielder and Goalkeeper datasets.

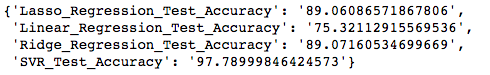
* 1. DATA MODELING

We use the PCA components to fit Linear and Polynomial Regression Models to the data and we use the scaled features to fit Ridge, Lasso and SVR regression methods. We perform K-fold cross validation with k=5 on the train data for each model. We find the optimal parameters using grid search function for Ridge and Lasso regression. The modified models are then trained on the training data and fit on the test data for prediction.

* 1. MODEL COMPARISON

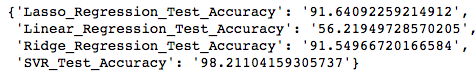
Shown below are the model test accuracy scores for each position:

**Defender:**

****

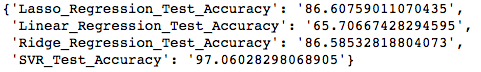
The best fit is SVR with highest test accuracy of 97.79%.

**Forward:**

****

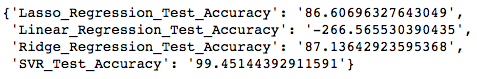
The best fit is SVR with highest test accuracy of 98.21%

**Midfielder:**



The best fit is SVR with highest test accuracy of 97.06%

**Goalkeeper:**



The best fit is SVR with highest test accuracy of 99.45%

* 1. PREDICTION

We find that SVR is the best fitting technique for all position datasets giving accuracy of 97% and higher. We use SVR to predict player rating.

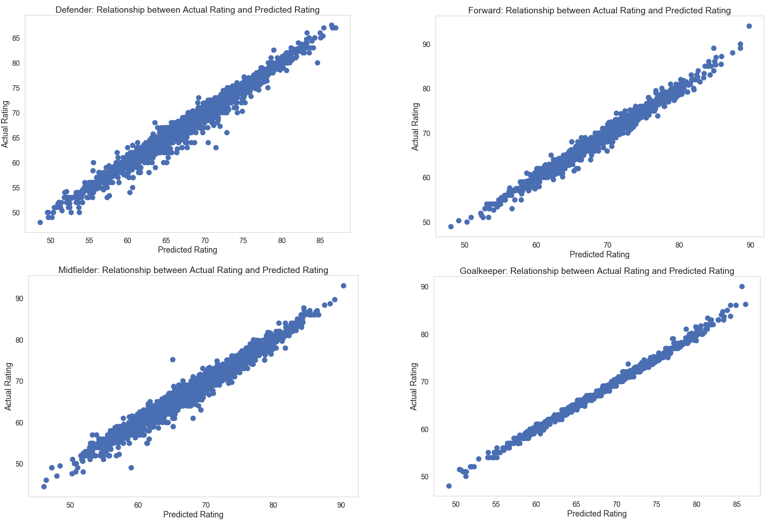


Fig 4.5.1

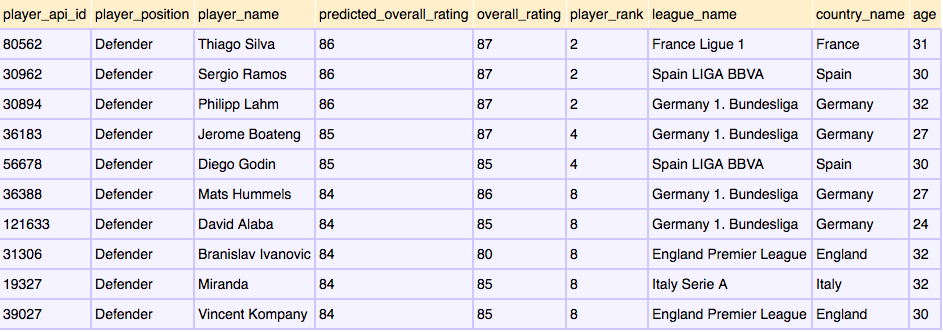
As we can see in the Fig 4.5.1, the scatter plots of the actual vs predicted ratings for all four positions show almost negligible dispersion. So, if the Actual is 50, the predicted will be reasonably close to 50 too. We can draw a regressed diagonal line through the data and the model will have a high R Square, since all the points would be close to this diagonal line.

1. RESULTS

The 'players\_with\_league' data has data of players who have played in the last two years, If a player does not have any match data since Jan 2013 then he is excluded from this table. We have merged our predicted datasets with this table in order to get a list of players who are active. This will help weed out non active players. However, the merge results in the loss of 26% (2771) players from the data.

**DEFENDERS:**

**TOP 10**

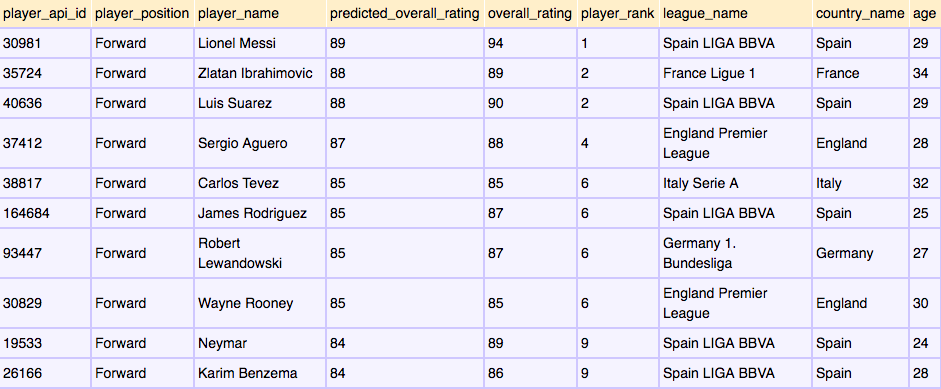
****

**SIGNIFICANT FEATURES**

Standing tackle, interceptions, marking, sliding tackle and reactions

**FORWARDS**

**TOP 10**

****

**SIGNIFICANT FEATURES**

**MIDFIELDERS**

**TOP 10**

**SIGNIFICANT FEATURES**

**GOALKEEPERS**

**TOP 10**

**SIGNIFICANT FEATURES**

1. LIMITATIONS

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

1. 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.

2. 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.

3. 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

4. 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.

1. FUTURE WORK

The FIFA 2017 data is now available, we can add that data to enhance and update the player ranking based on current data. With availability of more future data we can identify patterns in attributes of top players to then identify similar patterns in new players. This will help in scoping out up and coming players, this will help the client to make decisions related to player transfers as they could look at signing younger players with higher potentials at a lower cost.

We could also do additional exploratory analysis on effect of player nationality or age on players attributes. We can define more specific player positions as Left Back, Right Winger, Attacking Midfielder etc., to give more specific suggestions for player selection. Also, we can identify leagues or countries which produce top players. There is demand in identifying key performance attributes, effect of player’s associations with leagues and other such factors. Further studying of attribute differences between top players and bottom ranking players can help us understand attributes better.

If desired, we can further deep dive into analyzing the historical performance trend of the top players so that the client can make a final hiring decision considering the historical performance of the player.

1. CLIENT 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. While for goalkeepers, the trend is different with ratings remaining high from ages 25 onwards. So, 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. We were successful in building a model to predict overall ratings of Goalkeepers with 99% accuracy, Midfielders with 97% accuracy and lastly Defenders and Forwards with 98% accuracy. These models were based only on the numeric attributes available to us. By predicting the ratings for players by age the client can conveniently gauge the expected player performance for his upcoming matches