Predicting the best players for playing positions

The Problem: The client, a famous European Football Club wants to have a list of 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. The client requires us to build an algorithm that will predict the best 20 players for each specific position

The Solution:

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://www.football-data.co.uk/ : betting odds

http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. FIFA series and all FIFA assets property of EA Sports;

by Hugomathien and made available on Kaggle at https://www.kaggle.com/hugomathien/soccer

Cleaning and Transforming:

1. From the available database, we use the tables 'players' and 'player\_attributes' and identify each players preferred position of playing using their X, Y playing coordinates. We create a table of player\_positions to make merging of datasets easier

2. After merging all the available datasets, we keep only the features we need to predict the performance for a player. We use the feature 'overall\_rating' given by FIFA and we calculate the age of each player for the same rating date. This is to understand whether age of the player has impact on player performance. Alternatively, we can also collect rating for each player on the latest date available for the player and use that to correlate age and overall\_rating

3. We check this data for missing and incorrect values and clean it

Analysing:

1. Test normality of the data

2. Test the correlation and autocorrelation in data

3. Test linear relationship between overall rating and attributes

Modeling:

1. Modify the data so that each player has aggregated rating, attributes for each age of the player. Remove the last age row of each player as testing data.

2. With the remaining data as training dataset, use these aggregated attributes to predict the aggregated overall rating of the player.

3. Transform the data if necessary such that assumptions for different models are fulfilled. Fit regression models such as linear or polynomial regression, ridge regression, SVR and find the best fitting model with best parameters.

4. Use the best model to find the predicted scores from the test dataset.

5. Use these predicted scores to identify top 10 players in each position