Proposed Solutions:

We have 20+ player attributes from skill-based metrics such as long shots and dribbling to athleticism-based ones such as acceleration and jumping. So for feature selection, we will divide our data based on whether the player position is primarily involved in offense, defense, goalkeeping, or hybrid (and this can be done using nearest neighbor or k-means clustering), and run lasso regression to identify metrics important to each category of position, and combined the resulting sub-model from each cluster into one composite model.

As we will be attempting to forecast future performance as, vs extrapolating from the past, we can apply multiple decision tree models such as gradient-boosted decision tree to get a more accurate result that may be superior to the regression results we will obtain in the first stage. And we will use RMSE of the test data set to compare models and measure accuracy. Classification models such as naïve Bayes and random forest may also be incorporated into votes on a player’s valuation. Though neural networks may be suitable for our project, it may or may not be computationally feasible and we will explore this possibility. And then we can weigh the models by their RMSE, and build an ensemble model that may be more accurate than any single model alone.

In the end we will have one main model to determine if a player is undervalued or overvalued relative to his performance, and one model each for player performance and valuation for interpretability purposes.