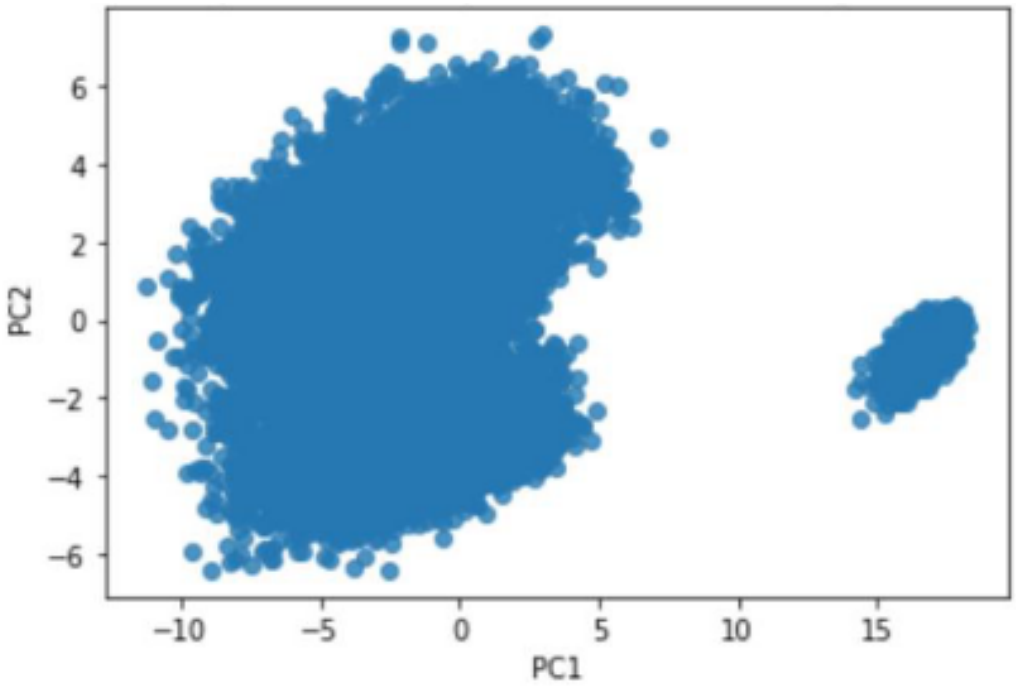


Figure 1: PCA Analysis

1 Context, Description of Data, and EDA

Football, or “soccer" as the US insists on calling it, is big. Top players earn tens or hundreds of millions of dollars per year, like Lionel Messi who’s taking home a paycheck of $92 million, and the European market alone is worth nearly €30 billion. A smart team manager, then, can earn a lot of money by putting together an effective team, and for that she needs to understand what a good player looks like. This requires significant amounts of data, and one of the best sources for those data is a franchise which in itself makes millions by accurately cataloguing and simulating footballers: EA Sports’s FIFA.

Our dataset consists of six years worth of player data scraped from the SoFIFA website, including player names, clubs, positions played, and dozens of skill rankings. It is mostly complete, although there were a couple of corrupted entries, but there were few enough of these that we could comfortably just drop them. More problematically, goalies and non-goalies were scored on completely different sets of skills, so no single row was complete. We solved this problem by filling in zeroes in empty cells, on the grounds that your average goalie probably has approximately zero skill level doing anything but being in goal, given the extent to which their jobs are different to any other footballer’s.

We performed a PCA analysis, shown in figure 1. While this very effectively separated goalies from non-goalies – the small clump on the right consists of all the goalies – it did not reveal much other structure, so we decided that further EDA would be more effective if it were guided by a known question and moved on to the first problem from the problem statement.

2 Problem 1: Predicting the Players

(Most of this material is taken directly from our milestone 3 report.)

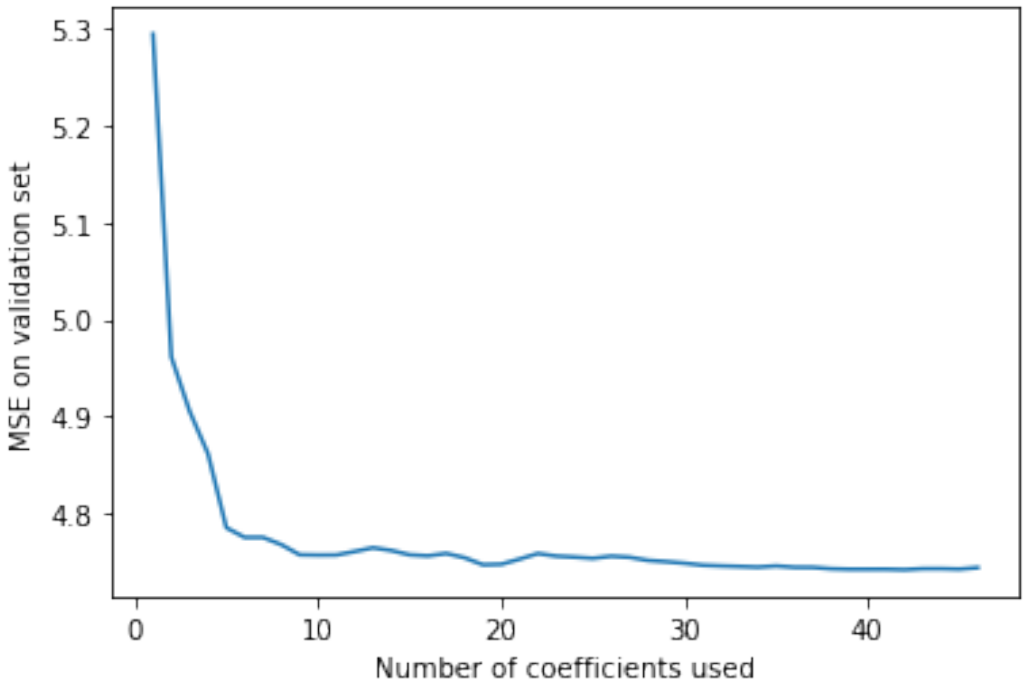


Figure 2: A very small number of components do most of the work

2.1 Problem Statement

Each football player in the dataset is given an overall score, calculated by domain scientists, which measures, out of 100, how good they are at playing football in general. Given statistics for players in 2019, including their overall scores in that year, we attempted to build a model to predict overall scores in 2020. Possible users for this model include scouts who want to know whether a good player will continue to play well.

2.2 Prediction

We took this problem as an opportunity to do some serious exploratory analysis on the dataset, in particular looking for correlations of any kind between the player's statistics in 2019 and his overall score in 2020. We found some extremely strong correlations, of which the most striking was a linear relationship between player score in 2019 and in 2020. There was also a slightly less obvious but still strong relationship between player potential and overall score. In other words, the equipment that SoFIFA had to rank footballers, whatever it was, was working: past performance plus potential did much of the work of future performance.

Since there were so many linear relationships, and since it would make sense for good footballers to get better as they trained, we decided to train a linear model. In particular, we trained a Lasso model, cross-validating to find the best parameter, in order to avoid overfitting and also to get an idea of which parameters were in fact important. We plotted the MSE versus number of components used and found that the returns from adding extra components diminished very quickly, as seen in figure 2.

Our model eventually achieved an MSE of about 4.5, far better than the k-NN model we used as a baseline, which never dropped below 10. Given the variation in individual regimens, club staff, injuries, and the myriad other things that can affect a footballer's ability to train, we believe that this

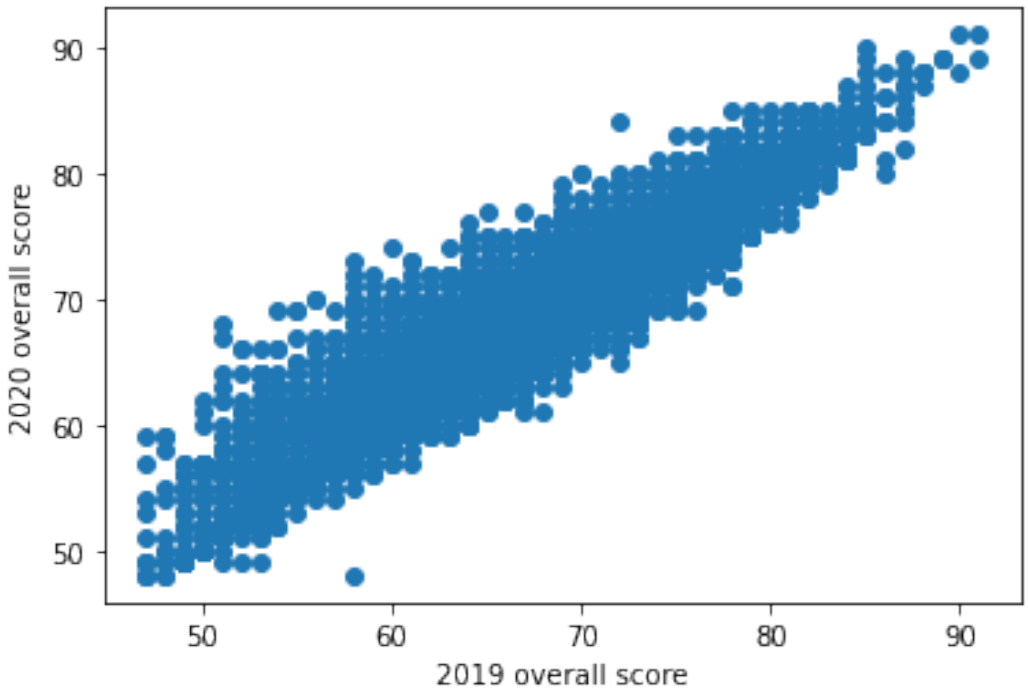


Figure 3: Overall scores in 2019 vs 2020 for players in the test set

is a pretty good result and one that we can definitely base problems 3 and 4, which require a baseline predicted overall score, on. However, it's still fundamentally a linear model, which means that any non-linear nuance in the dataset is lost on it. We do not believe there is enough of this nuance to worry us, but we could be wrong.

Another weakness that needs to be considered is that we don't know how the player overall score was evaluated. Perhaps understanding the the relation between the overall score and the other predictors could give us more insights on how to improve the model. However, the baseline model we have seems to be decent, with relatively low MSE scores.

2.3 Visualization

Graphically representing the "Overall" statistic for any given player doesn't tell us very much. It's one number, so a graphical representation of it is just that number written down. There's not much we could do with it, either, given that whether a footballer is any good doesn't really have any bearing on such easily-visualized things as what position he plays. Instead, we decided to make a very simple scatter plot, in order to get across the most crucial thing to know about the overall score, and the thing that underpins both our model in this problem and in problem 3: it doesn't change much between one year and the next. The result is figure 3.

3 Problem 2: Predicting the Positions

3.1 Problem statement

The dataset lists the positions, plural, that each player is comfortable playing. Given their statistics, and focusing on the data for players in 2019, we attempted to build a model to predict the positions

that each player could play. Possible users for this model include team managers who would like to know what newly-purchased players are capable of.

3.2 Prediction

"Predict a player's position" is not as simple as it sounds: players can play multiple positions, some of them don't play on the field at all during a normal game but are there as substitutes, and even the definition of a given position often depends on the manager as much as anything else. (One could imagine one manager who asked her center midfielder to be more aggressive, while another asked him to be more defensive.) Making this worse was the fact that the dataset contained fields for positions that a player could play, the position that he was currently playing, and the position, if any, that he played when he was last on a national team.

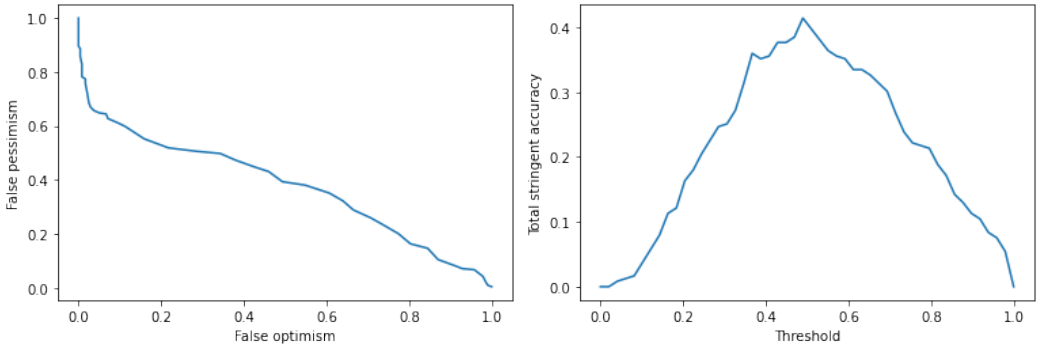
In the end, we made our choice based on the question of model utility. Predicting who plays what position on a national team isn't much use when you can just go and ask. Rather, the potential use case for our model would be a manager who had signed a player and wanted to know where he would best fit on the team. With this in mind, we decided to attempt to predict the list of positions that a given player could play, and to do this by training one model to predict the probability that he would be able to play each position.

Our PCA analysis from the previous part led us to expect very complicated decision boundaries, so we focused on decision tree models. (We could have also considered neural networks, but we had to train for fourteen positions at once, so we were forced to choose models which could be trained more quickly.) Since there is no one perfect way to be the best striker or the best left wing-back, a model which tried many different approaches to classification and aggregated the results seemed appropriate, so we settled on a random forest, constraining the maximum depth of each tree to avoid overfitting.

Fourteen random forests later, we could start making predictions and scoring the model. Here arose further difficulties. Simple classification error would not work here, as we would not want our model to be marked down for listing positions in the wrong order. Instead, we developed the concept of "false optimism," when the model predicts that a player will play in at least one position in which he does not actually play, and the similar "false pessimism," when the model fails to predict that a player plays in a position which he actually does play. We wrote functions to calculate the fraction of players for whom the model was falsely pessimistic, falsely optimistic, or, as happened in a some cases, both. We also defined "stringent accuracy": the model is stringently accurate for a player if it is neither falsely pessimistic nor falsely optimistic.

The trade-off between false pessimism and false optimism gave rise to the question of where to set the predictive threshold. Does a player count as a center midfielder if the CM model outputs 0.5? How about 0.3 or 0.7? By varying the threshold from 0 to 1, we were able to graph false optimism against false pessimism and produce a variant of ROC curve, as in figure 4a. We also plotted stringent accuracy against chosen threshold, as in figure 4b. These plots are for the test set; those for the train set are less interesting, having the optimism-pessimism curve take on much more of an L shape and the accuracy curve look like a gentle hill with a peak around 0.5. Since the peak was there, we decided not to risk overfitting by choosing a more optimal threshold for the training set and instead left the threshold at the default 0.5.

The overall stringent accuracy tops out at about 0.4, which does not seem very good at first glance. However, note that there are fourteen models involved, and all fourteen of them have to be correct. For a stringent accuracy of 0.5 exactly, we would expect the models to have average accuracy of $\sqrt[14]{0.5} \approx 0.95$. As is, the models we trained averaged accuracy of about 0.92 on the test



(a) False optimism-false pessimism curve

(b) Accuracy versus threshold

Figure 4: Model metrics for the test set

set, and $0.92^{14} \approx 0.31$, so the aggregate strategy is actually doing much better than expected. We believe this to be due to correlation between certain positions: good center midfielders often make good center attacking or center defending midfielders.

We also calculated AUC scores for each individual model on the test set, and got a mean of ≈ 0.90.

These are very good numbers and we are happy with our model's performance, but we must keep in mind that it is not as useful as possible for an actual manager because it is not in any way interpretable. Reading tall decision trees is difficult at the best of times, and when they are combined in a random forest it becomes impossible. We therefore recommend that our model be used less as a way of placing players on a pitch and more as a way of finding untapped potential that might be worth exploring in real life.

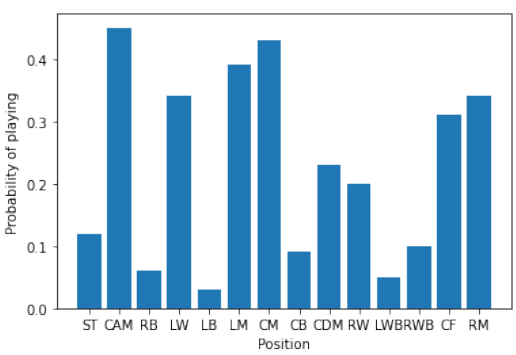
3.3 Visualization

While the "internal" visualizations that we produced to determine the effectiveness of our model could be quite basic, the "public-facing" ones, designed for end-user consumption, would have to be more complicated and easier to read. We could certainly have developed visualizations for certain properties of the dataset in aggregate – which positions were the most popular, which predictions were hardest for our models to predict, and so on – but that would go against the main purpose of our model, which was to take an individual player or set thereof and give a manager a sense of where he would play well, so we did not investigate that. We didn't investigate visualizing anything for goalies either: since we can identify them with 100% accuracy, the visualization for a goalie is a box with the word "Goalie" in large text.

Instead, we developed two types of visualization. The first, the most basic possible, was a bar chart giving every possible non-goalie position and how likely a player is to play in each. Our guinea pig, by virtue of being the first non-goalie member of the test set when ordered by who came first in the CSV file, was Kevin De Bruyne, and figure 5a shows his graph.

This is boring to look at and, worse, doesn't give an immediate idea of De Bruyne's capabilities, so we developed an alternative, shown in figure 5b. This makes it easy to tell at a glance that he is a versatile attacking midfielder who can even, in a pinch, play forward or striker. According to a 2018 Premier League profile, this is in fact the case.[1]

[1] `https://www.premierleague.com/news/686309`



(a) Predicted positions

(b) Visualized pitch

Figure 5: Predictions for Kevin De Bruyne

4 Problem 3: Club with the Best Staff

4.1 Problem statement

Players play football in clubs, and clubs have a great influence over the development of a player. We attempted to develop a way of ranking clubs out of 100 in terms of which are best for a player's professional development, based on the change over time in the average statistics of players in that club. Possible users for this model include players trying to decide if they should accept a transfer offer.

4.2 Prediction

For this part, we first needed to figure out which clubs played in each of the five leagues, we also needed to make sure the comparison is fair, so we only picked the clubs that played in each of the past five years. We end up with 50 clubs total.

After picking these clubs, we cleaned up the data in a similar fashion to what we did in the previous parts. Since we only want to analyze how the player stats changed over the years, we removed all the data points that do not directly correspond to player skills. We end up with 81 predictors total.

After that we normalized the data, which proved helpful when running the model and predicting future stats. Using the normalized data, we took the average of each skill for each club, and for each year. We end up with five data frames, each corresponding to a year and containing info about the clubs.

We visualized some of the change of the skills over the years in order to see if we could see any patterns, which led us to decide on a linear regression model, because there was not much change in the average skill stats. So, we ran a linear regression model for each skill at each club, and recorded their slopes which we used to determine the improvement score for each club. We did that by taking the average of all the slopes, a positive average means there was an overall improvement, negative average means the club stats became worse, and zero average means there was no overall change.

We normalized these averages to score the clubs out of 100, then we sorted them out in descending order.

It is important to note that although we can notice the differences in scores between the clubs, these differences are not significantly different, but relative to one another, we can see this pattern.

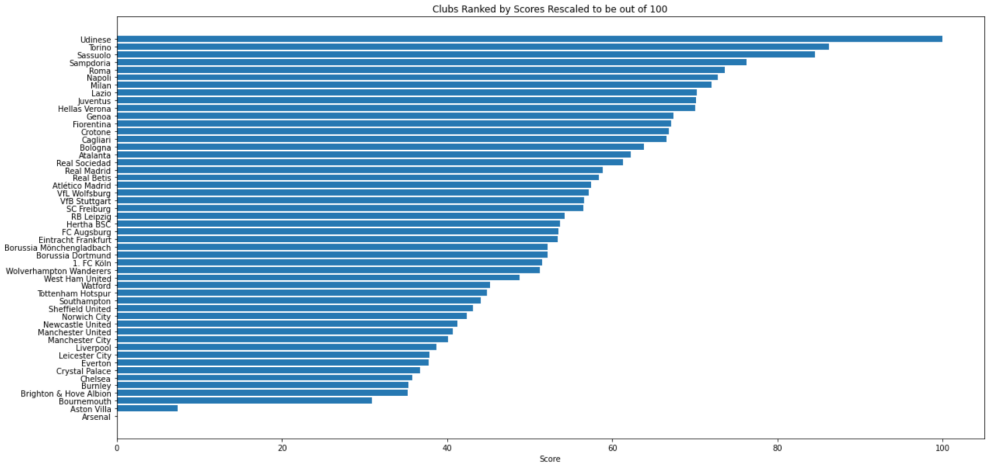


Figure 6: Clubs Ranked by Scores out of 100

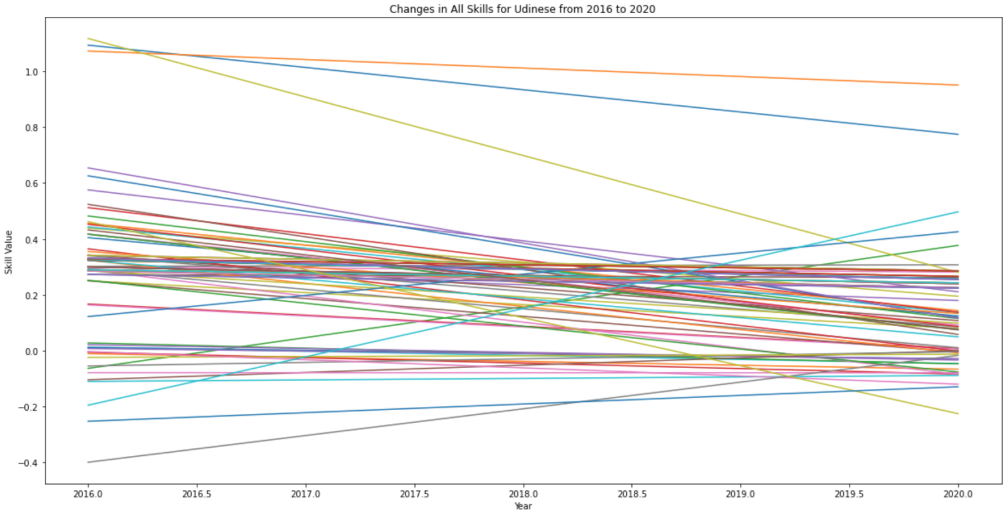


Figure 7: Changes in Skills for Udinese

Figure 7 shows the changes in skills for the club that ranked first. Although it looks messy, it gives us a good idea on how the overall club performance relates to the individual skill stats. Many of the lines look flat, some have negative slopes, and the rest have positive slopes. We saw a very similar pattern when plotting the skill changes for other clubs. This explains why the average slopes have relatively low values (e.g. 0.04).

A main weakness for this approach is that we relied on club averages to rank the club performances, which is reductive and loses information. For example, a club that has good goalies and bad strikers, is equivalent to a club with with bad goalies and good strikers. So, this method is not very reliable when predicting which club is better for improving players' stats. A better approach, which we utilized in problem 4, is to model the stats of each player individually, and only then average them to determine the overall club performance change.

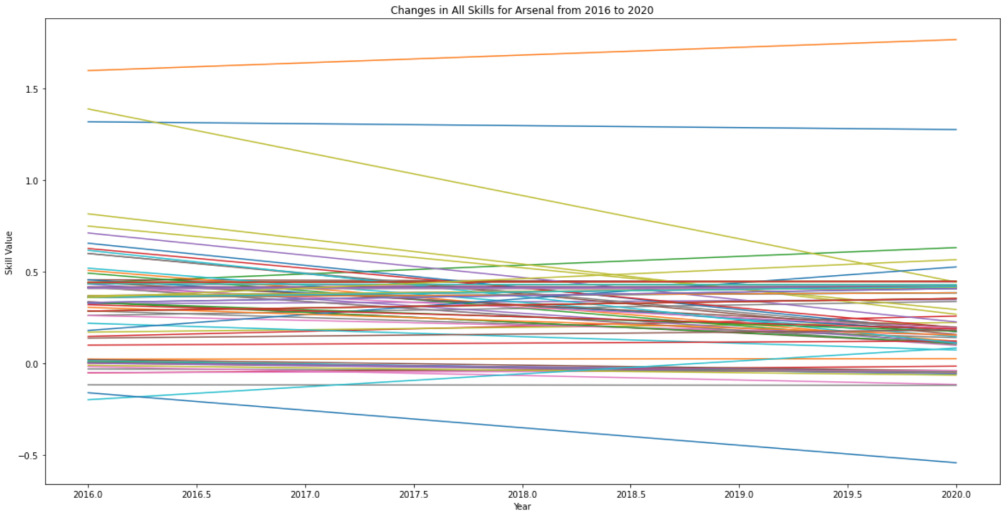


Figure 8: Changes in Skills for Arsenal

5 Problem 4: Skill Stats for 2020-2021 Season

5.1 Problem statement

As in problem 1, we attempt to build a model to predict players' scores, except this time we attempt to predict all skill scores, rather than just overall scores. Possible users for this model include scouts and mangers who would like to know what sort of person a player might develop into.

5.2 Prediction

This part is the most difficult, because there is very little data for a given player and a given statistic on its own. We get at most five data points, which is enough for a linear regression, possibly, but not enough for anything else. With this in mind, we decided to aggregate those players who had similar training regimens, similar management, and similar successes or failures. We grouped by club.

To do this, we first took the roster for each club in each year, calculated the mean statistics for that roster, calculated the mean statistics for that same roster in the next year – which involved looking them up in a list of players, because any given member of the roster might have changed club – and thereby calculated the improvement that an average player in that club experienced in that year. This was very similar to what we did in problem 3, and that similarity will become even more pronounced in a moment.

However, clubs are only half the battle. A good player will do fairly well even in a bad club. For this reason, to predict the statistics for a given player, we started by calculating the amount by which his statistics had improved in a given past year. We call this the "player delta" δp. We compared this to the "club delta" δc, the amount that his club's mean statistics had changed over that year. Then, to quantify the extent to which the player went above and beyond his club's training regimen, we define the "overachiever factor" $o = \delta p - \delta c$. (There is no reason why $o > 0$. We call players for whom it isn't "underachievers" in that statistic.)

Since we had, at best, four data points for o, we felt it unwise to fit a regression to it at all, and instead took the mean $\bar{o}$. This is, given the small amount of data we had, a good representation of the amount that, left on his own, a given player would improve from 2020 to 2021.

The final piece of the puzzle was to factor in the clubs. This was harder than it looked. Many clubs did not play in the same leagues for all five years, and in problem 3 we had restricted ourselves to only scoring clubs which had. (This seemed like a good idea, on the grounds that moving from league to league can lead to large shake-ups in club management.) In short, we couldn't meaningfully predict the extent to which many clubs would change from 2020 to 2021. We could have fit linear models even to clubs that jumped between leagues, but, again, that might well have been unwise and irresponsible. Confronted with this classic missing data problem, we did the best we could, and, given that the exact changes in most mean club statistics were very small year over year, we simply assumed that clubs we did not score in part 3 did not change statistics.

For those clubs that we did score in part 3, we extended the linear regressions calculated to make the scores to predict the new statistics in 2021. As our MSEs from those regressions were quite low, we felt justified in using them rather than fitting more complicated models to our small datasets. From this prediction, we could calculate δc for each club. (This δc is also equal to the slope of the linear regression line, but predicting the exact club scores in 2021 fit better with the way we arranged, and internally thought of, our code.) Finally, we took each player's raw statistics in 2020, added the club delta if available (if it wasn't available, we did nothing because we were assuming it was 0), and finally added the player's $\bar{o}$.

The most difficult part of this problem was drawing the line between what we could meaningfully fit a regression to and what we couldn't fit any parametric model to at all. We eventually decided that club statistics provided just enough, but no fewer, and by the fence-post principle (a fence post with n panels has $n+1$ posts), we therefore did not have enough to fit a model to o. Furthermore, o lends itself more to averaging than club statistics: while a club might trend upwards or downwards in a given year due to budget, morale, or leadership, o is more a measure of a player's commitment, will to win, and ability to work harder than his compatriots, and these interior qualities are harder to consciously develop than, say, good sports nutrition or making sure your players get enough sleep. They are therefore more likely to change less over time, or to change more randomly if they do.

Our model is very limited, relying as it does on data going back only a few years and predicting as it does each statistic independently. If we had had more time or more data, or preferably both, we would have liked to look into correlations between statistics that would enable us to get a better idea of exactly what the club and player deltas meant. For instance, if a player suddenly improved on passing and taking free kicks, but not on interceptions, a more advanced model would be able to understand that he had been training specifically for accuracy. However, our model does make good sense, it is easy to train and deploy, it is highly interpretable, and it nicely separates the work of the player from the work of the club. While we would not make it public without a warning label explaining exactly what it did and the many simplifying assumptions it made, we believe that, at least as a baseline for further exploration, it is perfectly workable.

It would also be interesting to combine this model with the model from problem 2, to predict whether a player's positions would change. Possible users for this combined model would include managers who wanted to plan for what sorts of players to buy a few years in the future.

In conclusion, we have pretty good models to not describe the data, but to also predict the future performances of clubs as well as individual players. The data set needed a lot of cleaning, and the missing data points for many of the players posed a lot of challenges. Perhaps, a more 'complete' data set would be even more reliable to make predictions.