MIT MANAGEMENT SLOAN SCHOOL

Master of Business Analytics

15.095: Machine Learning Under a Modern Optimization Lens

Soccer players market value prediction and optimal transfer strategy prescription

Oscar COURBIT

Andrea ZANON

github.com/azanon00/MachineLearning_Project

December 8, 2022

Contents

1 Introduction

The transfer market is the best opportunity for soccer clubs can rely on to improve their team and compete for a better placement in the following year. However, identifying the right transfer value (the price for which a team is willing to buy/sell a player) is complicated since it depends on many factors. It is usually a reflection of what the player is worth in the future instead of how well they did in the past (often the two are correlated, but not necessarily). Moreover, even when the right transfer value is considered, teams have many options to decide which players to buy/sell, and oftentimes this decision falls on prestige rather than truly assessing how well the player would fit in the team. As a first step, we plan to use Machine Learning to predict the future transfer value of a player. Once we have that information, we can put ourselves in the shoes of a football team and then propose a prescription model to decide which players to buy and which players to sell to minimize the net spend while improving the (future) quality of the team.

2 Data

2.1 Dataset

To tackle this problem, we have access to two sources of data:

- The yearly FIFA's players rating datasets between 2014-2015 and 2021-2022, including market value, personal details (nationality, height, weight, age, preferred foot, ...), performance statistics (overall rate, pace, passing, defending, dribbling, ...) and contract details (wage, release clause, contract duration, ...). For instance, the 2021-2022 dataset has approximately 20,000 observations and 110 features in total (many of which are categorical).
- A dataset gathering all the summer transfers from 1992/93 up to the last transfer window (summer 2022), that includes each players' personal details, origin (club, country), destination (club, country), and type of transfer (loan or not). For example, the summer 2022 dataset has approximately 2,000 observations and 12 features.

Because of many other economic factors and of the evolution of the players' performances, predicting market values over multiple soccer seasons would be overly complex and might not give accurate results. Therefore, we decided to focus on the last three full soccer seasons (2019-2020, 2020-2021, 2021-2022) and, for each season, we joined each corresponding rating and transfer dataset.

2.2 Pre-processing

We decided to translate players' specific role (there are more than 20) to four main roles: attacker, midfielder, defender and goalkeeper. A few additional columns were converted to categories (preferred foot, attacking work rate, defending work rate) to keep more valuable information; we used either one-hot encoding or ordinal encoding to translate these features.

We then split the data into a training/validation set and a testing set: our training set includes the 2019-2020 and the 2020-2021 data; our testing set corresponds to the 2021-2022 dataset (as we will focus on the 2022 transfer period in the prescription part).
Overall, the training/validation set has approximately 40,000 datapoints, and the test set 20,000.

2.3 Missing data imputation

After the data-preprocessing, we realized that some specific data was missing for all goalkeepers. Deleting the corresponding rows was not an option, first because it would have introduced bias in the prediction, but also and foremost because this would have excluded a whole category of players from being considered as possible transfers. To tackle this issue, we implemented optimal missing data imputation using `IAI.ImputationLearner(:opt_tree)`.

3 Market Values Prediction

High-quality predictions are extremely important in this setting: knowing precisely what the true value of a player is allows to plan ahead the market strategy and to understand how to negotiate with other clubs.

3.1 Models implemented

We decided to try many different regression algorithms studied in class: Linear Regression, Robust Regression (LASSO), Random Forest, XGBoost, Holistic Regression, Optimal Regression Trees (ORT), Optimal Regression Trees with Hyperplane Splits (ORT-H), Optimal Regression Trees with Linear Regression on the leaves (ORT-L).

All the models were trained on training/validation using a grid search and 5-fold cross-validation to select the best combination of hyperparameters.

3.2 Prediction Results & Model Choice

Model	**Runtime** (s)	**R2**	**RMSE**	**MAE**	**MAPE**
Linear	**0.5**	70.2	4.16	1.91	126.1
Lasso	0.8	66.4	4.54	1.57	89.7
Random Forest	86	87.5	2.75	0.58	21.2
XGBoost	128	**88.7**	**2.60**	**0.51**	**19.2**
Holistic	4800	71.5	4.14	1.73	114.7
ORT	373	80.4	3.37	0.96	67.1
ORT-H	49380	71.7	3.86	0.85	54.8
ORT-L	3260	73.3	3.99	1.58	91.3

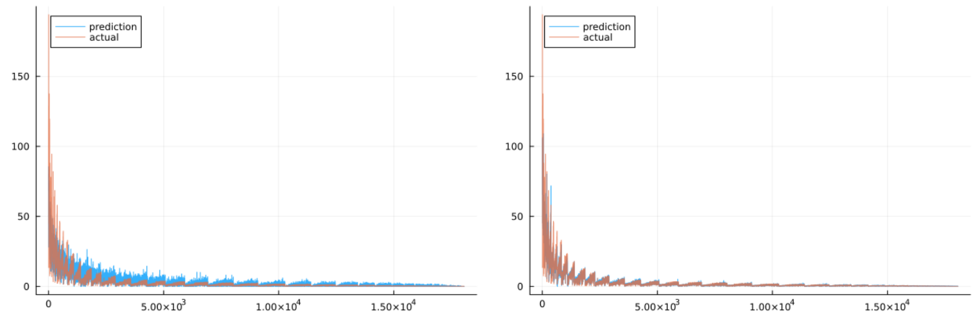
Table 1: Out-of-sample Results

The above table (Tab.1 shows that the most accurate models are not very interpretable (XGBoost, Random Forest). In our task, performance is fundamental since it is incredibly important to come up with an accurate economic prediction of the market value in order to plan well the transfer campaign. As a result, our best model choice would be **XGBoost** since it outperforms the other models in all the metrics evaluated, and is reasonably fast. As we can also see from the figure below (Fig.1), XGBoost predictions are very accurate, compared to a Linear Regression (that we will use as a baseline prediction).

3.3 More interpretability

While it is true that predictive performance is mostly what we care about in this task as it is necessary for the downstream prescription, there is no denying that interpretability can matter as well, to give a soccer manager the tools to understand the decisions made by our prediction models. For instance, if we consider a very young player that is worth a lot, a manager is definitely interested in understanding why.

In our case, XGBoost has the strongest performance, but this algorithm seems like a black box. Therefore, we proceeded as follows: first, we selected the most significant features on which



(a) Linear Regression prediction

(b) XGBoost prediction

Figure 1: Out-of-sample Predictions v. Real market values distributions

XGBoost relies (`XGB_model.feature_importances_`); then, we implemented an Optimal Regression Tree on these variables to predict **the predictions made by XGBoost**. We obtained the following tree (Fig.2).

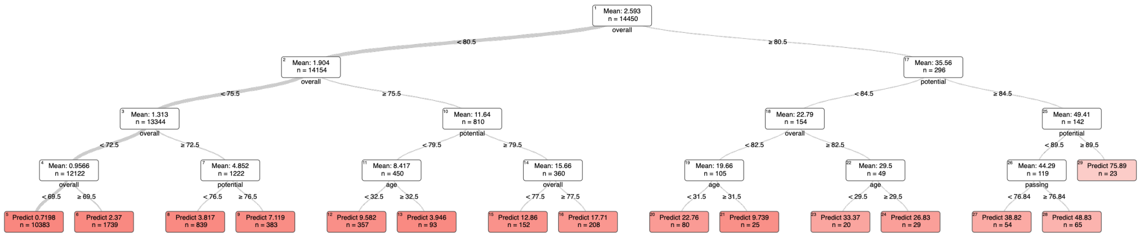


Figure 2: ORT that predicts the XGBoost output; $\mathrm{R}^2 = 0.92$

This Optimal Tree is fairly precise ($\mathrm{R}^2 = 92\%$): it means that this tree can accurately represent the decision process made by our XGBoost algorithm to predict market values. In this way, a soccer manager can rely on this "close-to-human-thought-decision-process" tree to understand how XGBoost works. It is clear that the features `overall` and `potential` dominate the predictions. However, we believe it is also valuable to understand which features were more valuable if we removed one or both of them. Results are summarized in the appendix: we considered one tree without these two features (Fig.4, $\mathrm{R}^2 = 71\%$) and one tree where we only removed the `overall` feature (Fig. 5, $\mathrm{R}^2 = 86\%$). We can note that the latter is also precise and involves a greater diversity of features, which can give more explainability to XGBoost's decisions.

4 Transfer Strategy Prescription

4.1 Problem Formulation

Idea: minimize, for a given club, the budget spent during the transfer period, subject to the constraint of achieving a given δ in team rating improvement.

$$\begin{aligned} \min \quad & \text{net\_spend } (\iff \text{ max} \quad \text{profit}) \\ \text{s.t.} \quad & \text{improvement} \geq \delta \\ & \text{other constraints} \end{aligned} \tag{1}$$

Solving this problem, we are putting ourselves in the shoes of some specific clubs, understanding what their optimal transfer strategy is. We can then compute the results as average performance across the different teams.

4.1.1 Parameters

For every player:

- m_i: linear combination of the overall rate and the future potential rate of player i
- a_i: age of player i
- $\mathbf{x}_i$: some other data about player i (current rates, remaining years of contract, release clause, wage...) – the independent variables in the prediction model
- v_i: market value of player i next summer – the dependent variable in the prediction model

We will use the notation ^{o} to refer to players outside of the club, and ^{c} for players who are part of the club.

For the given club:

- δ: by how much we want to improve the overall rate of the club at least
- γ: by how much the average age can change at most
- n: total number of players who are currently outside of the club
- k: total number of players who are currently part of the club
- $i = 1, ..., n$: players that can be recruited (outside of the club)
- $j = 1, ..., k$: players that can be sold (inside the club)
- N_{min}, N_{max}: minimum and maximum number of players that can be allowed
- B_{max}, S_{max}: maximum number of players that the club can buy and sell (to do not modify too drastically the club's structure and identity)
- $\mathcal{A}_c$, $\mathcal{M}_c$, $\mathcal{D}_c$, $\mathcal{G}_c$, $\mathcal{A}_o$, $\mathcal{M}_o$, $\mathcal{D}_o$, $\mathcal{G}_o$: sets of attackers, midfielders, defenders and goalkeepers inside and outside the club.
- N_c^a, N_c^m, N_c^d, N_c^g, N_o^a, N_o^m, N_o^d, N_o^g: number of players in each of these previous sets
- N_{min}^a, N_{min}^m, N_{min}^d, N_{min}^g, N_{max}^a, N_{max}^m, N_{max}^d, N_{max}^g: minimum and maximum number of attackers, midfielders, defenders and goalkeepers that can be tranfered

While a few of these parameters have been arbitrarily defined, the majority have been computed from past observations. Generally speaking, these parameters can easily be changed before running the optimization in order to reflect the team's needs.

4.1.2 Decision Variables

- $b_i = \begin{cases} 1 & \text{if we decide to buy player i, i = 1,...,N} \\ 0 & \text{otherwise} \end{cases}$
- $s_j = \begin{cases} 1 & \text{if we decide to sell player j, j = 1,...,k} \\ 0 & \text{otherwise} \end{cases}$

4.1.3 Objective Function

The goal is to minimize the budget spent during the transfer window, which means to maximize the profit:

$$\max_{\mathbf{b},\mathbf{s}} \mathbb{E}_{\mathbf{v}^o,\mathbf{v}^c}\left[\sum_{j=1}^{k} s_j v_j^c - \sum_{i=1}^{n} b_i v_i^o \,\middle|\, \mathbf{x}^o, \mathbf{x}^c\right] \tag{2}$$

- $R(s_j, v_j^c) = s_j v_j^c$: the revenue made by the sale of player j
- $C(b_i, v_i^o) = b_i, v_i^o$: the cost of recruiting player i

4.1.4 Constraints

The most important constraint is the **improvement constraint**, which can be written as follows:

$$\sum_{j=1}^{k}(1 - s_j)m_j^c + \sum_{i=1}^{n} b_i m_i^o \;\geq\; (\frac{1}{k}\sum_{j=1}^{k} m_j^c + \delta) \cdot (k + \sum_{i=1}^{n} b_i - \sum_{j=1}^{k} s_j) \tag{3}$$

It means that the difference between the new average overall rate of the team and the previous one has to be greater than a given δ.

Moreover, we have additional constraints:

- $\sum_{j=1}^{k}(1 - s_j)a_j^c + \sum_{i=1}^{n} b_i a_i^o \;\geq\; (\frac{1}{k}\sum_{j=1}^{k} a_j^c + \gamma) \cdot (k + \sum_{i=1}^{n} b_i - \sum_{j=1}^{k} s_j)$
  Average age change allowed is determined by γ
- $\sum_{i=1}^{n} b_i \leq B_{max}$
  Limit to the number of recruits
- $\sum_{j=1}^{k} s_j \leq S_{max}$
  Limit to the number of sales
- $N_{min} \leq N_c + \sum_{i=1}^{n} b_i - \sum_{j=1}^{k} s_j \leq N_{max}$
  Number of players in the team must be within a certain range
- $N_{min}^{pos} \leq \sum_{i \in pos_o} b_i - \sum_{j \in pos_c} s_j \leq N_{max}^{pos} \quad \forall pos \in \{a, m, d, g\}$
  For each position, we can buy/sell a limited number of players
- $b_i, s_j \in \{0, 1\}, \forall i = 1, ..., n, \forall j = 1, ..., k$
  Binary decision variables

Some of these constraints were included in the formulation since the beginning, others were necessary because of the results we obtained. For instance, when analyzing our first results, we observed that the optimal strategy delivered by our algorithms was to recruit mainly attackers or goalkeepers: thus, we had to balance our recruiting strategy. Another issue was that our optimal strategy seemed to push us to buy cheap older players and sell younger ones; this is not coherent with what happens in real life: thus, we added the age constraint to keep a smooth dynamic transition after the transfers period.

While it would also be possible to enforce other constraints, such as a budget constraint (note that a Robust Optimization formulation would be needed to account for the uncertainty in the predictions) or the impossibility to buy/sell certain players, we chose to keep the formulation simple but realistic.

4.2 Prescription task

The problem For a given club, suppose that we have data $(\mathbf{x}_i^o, v_i, b_i)$, $i = 1, \dots, \tilde{n}$ and data $(\mathbf{x}_j^c, v_j, b_j)$, $j = 1, \dots, \tilde{k}$, which are players' information and market values from previous years. We will note these as $(\mathbf{x}_l, v_l), l = 1, \dots, \tilde{n} + \tilde{k}$.

We now have a new data point $\mathbf{x} = (\mathbf{x^o}, \mathbf{x^c})$ corresponding to the same club trying to recruit and sell during a new transfers period. We can use predictive models, train them on these previous transfer periods, and use them to predict an output $v = (v^o, v^c)$ when given a new input $\mathbf{x}$.

Regress and Compare Let's suppose that we have a machine learning method which gives accurate predictions: $f(\mathbf{x}) = \hat{v} \approx \mathbb{E}[v|\mathbf{x}]$. Then we can plug in our prediction $\hat{v}$ for the unknown outcomes v, and rewrite the objective function (while keeping the same constraints) as:

$$\max_{\mathbf{b},\mathbf{s}} \sum_{j=1}^{k} s_j \hat{v}_j^c - \sum_{i=1}^{n} b_i \hat{v}_i^o \tag{4}$$

While at the beginning of the project we imagined trying a variety of different prescription techniques (KNN, CART, Random Forests), because our objective function is linear, we would not have better results using a weighted scheme. Indeed:

$$\mathbb{E}_{\mathbf{v}^o,\mathbf{v}^c}\left[\sum_{j=1}^{k} s_j v_j^c - \sum_{i=1}^{n} b_i v_i^o \,\middle|\, \mathbf{x}^o, \mathbf{x}^c\right] = \sum_{j=1}^{k} s_j \, \mathbb{E}_{\mathbf{v}^o,\mathbf{v}^c}[v_j^c | \, \mathbf{x}^o, \mathbf{x}^c] - \sum_{i=1}^{n} b_i \, \mathbb{E}_{\mathbf{v}^o,\mathbf{v}^c}[v_i^o | \, \mathbf{x}^o, \mathbf{x}^c]$$

And these expectations are exactly what is predicted by the machine learning method we use. Therefore, we sticked to the Regress and Compare method and focus on getting the best possible predictions.

4.3 Results

We run the aforementioned prescription framework for different teams of different quality. The optimization is run with different values for the parameter δ (the minimum change in average quality we require to obtain) so that a curve is obtained, describing the necessary revenue (or cost, if negative) to achieve the given δ improvement. Intuitively, the higher the improvement, the lower the revenue/profit.

One important thing to remember is that we are making prescriptions for the 2021-2022 season, which is covered by our test set. Therefore, the results are truly out-of-sample.

For each team, we plotted three different results (Fig.3):

- A blue curve, representing the optimal strategy with a complete information of the true market values of all players
- An orange curve, representing the optimal transfer strategy given our market values predictions (Linear Regression, or XGBoost).
- A green curve, representing the optimal transfer strategy given the players we would buy using our predictions, but computing the profit using their true market value. This is the right curve to study in practice.
- Vertical and horizontal lines represent respectively the real-life net spend and improvement obtained in the last transfer window.

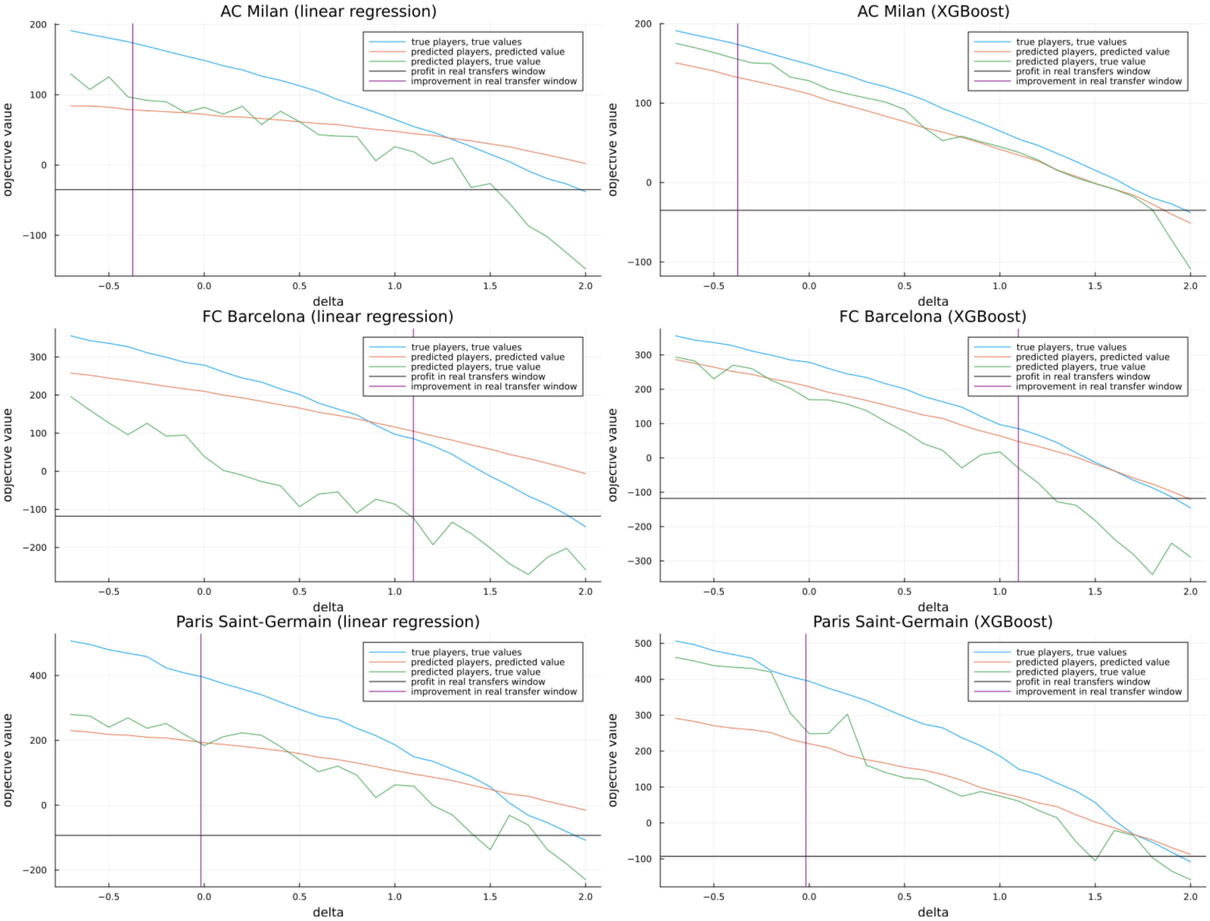


Figure 3: Prescription results relying on Linear Regression predictions (left) and XGBoost predictions (right)

Using XGBoost predictions, we can see how the blue curve and the green curve are much closer than using Linear Regression, meaning that we are obtaining more truthful results.

Moreover, Using the true 2022 transfer data, for each of these clubs, we can compute their profit and improvement at the end of this transfer window. Then, thanks to these plots, we can compare the real-world results to the profit and improvement obtained by our prescription model.

These results are summarized bellow (Tab.2).

Instance v. Club	AC Milan		FC Barcelona		PSG	
	Budget	Delta	Budget	Delta	Budget	Delta
2022 real transfers period	-35 M	-0.38	-118 M	1.10	-93 M	-0.02
Prescription, same budget	-35 M	**1.80**	-118 M	**1.28**	-93 M	**1.47**
Prescription, same improvement	**150 M**	-0.38	**-33 M**	1.10	**261 M**	-0.02

Table 2: Improvements achieved by our prescription method based on XGBoost predictions, compared to the 2022 transfers period

We can clearly see that optimal prescriptions have an edge over the market strategy followed by each team. Of course, it must be considered that reality is more complex and many aspects are evaluated when deciding whether to buying/selling a player. However, we are confident that by following the same approach in a more realistic scenario would be highly beneficial.

Let's also observe that both AC Milan and PSG have a negative revenue balance while not improving their team ($\delta < 0$). PSG even spent around $100 million dollars for a zero δ overall improvement, but this can have some explanation: PSG, owned by the Qataris, did not hide that their recent strategy was to attract renowned players (sometimes overrated) to achieve a great marketing blow for the World Cup in Qatar. On the contrary, Barcelona recently faced serious economic difficulties: this might be why their strategy seems very close to a "Moneyball" optimal strategy, just like the one we built (attracting cheaper, underrated and promising players).

5 Conclusion and possible improvements

To conclude, through this project it is clear that better predictions lead to better prescription, and by adopting an optimal strategy teams can have a significant edge.
Of course, reality is multifaceted and much more complex, and when devising their transfer strategy, clubs consider many more different factors than just the sheer economic cost of the player. Still, we strongly believe that by adapting this work to a real scenario (which means, imposing constraints to make the situation coherent with the overall strategy of the specific club), teams can have significant benefits.

In particular, a direction of improvement that is certainly worthwhile to explore is taking into consideration not only the transfer value of the player, but also his release clause, wage, and remaining length of the contract. In addition, teams can sign free transfers or loan players. One other layer of complexity is that different teams are not equally "popular". All these considerations would be very useful and would highlight the existence of a relatively significant **uncertainty** around players' market values: an important next step would be to develop a **Robust Optimization** model.

While all these aspects would certainly lead to a better model, we are confident that our analysis is realistic and meaningful.

Individual contributions

Throughout the project, contributions of both team members have had equal weight. Since the beginning we organized periodic events to define future steps, split work and assess progress. Moreover, we constantly kept each other informed when implementing new parts of the code, asking for feedback, and double-checking the work done.

The project can be broken down into four main parts (data pre-processing, prediction, prescription, communicating results through report and presentation), what follows is individual contribution in each of them:

- **Data Pre-processing:** Andrea did mostly the data cleaning and feature encoding, while Oscar implemented the missing data imputation and merged the datasets.
- **Prediction:** Andrea put the foundation for most of the models, and Oscar finalized all the models and run experiments.
- **Prescription:** Oscar defined the core of the mathematical formulation; Andrea finalized details, coded it in Julia, and run experiments.
- **Presentation of the results:** We both participated in an equivalent way, developing more specifically the axes of our own work, but also checking and improving the work of our partner.

As we mentioned at the beginning, we insist on the fact that both of us took part in everything, so the distinction is often blurred. This project was a collective construction from the beginning to the end, thanks to an active, enthusiastic, and continuous collaboration

Appendix

Interpretable trees

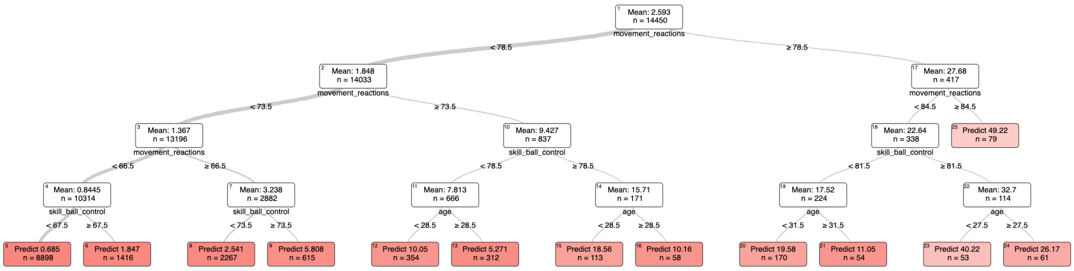


Figure 4: ORT that predicts the XGBoost output; $\mathrm{R}^2 = 0.71$

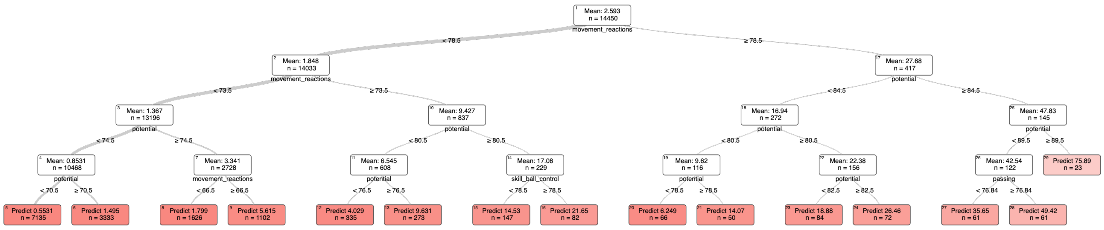


Figure 5: ORT that predicts the XGBoost output; $\mathrm{R}^2 = 0.86$