Master of Business Analytics

15.095: Machine Learning Under a Modern Optimization Lens

Interpretable end-to-end player prescription via machine learning and robust optimization

Marco Antonioli
Tom Wright

January 15, 2023

Contents

1 Introduction

Every year, football teams have the chance to modify their rosters by acquiring or selling players. This process happens twice a year: once between two seasons, during summer, and once in the middle of the season, during winter. The summer transfer window is usually the moment in which teams want to improve their roster. During the winter session however, teams want to keep synergies built throughout the season and usually only adapt to the condition of their players; for example acquiring players for positions in which they have injuries.

We are going to position ourselves as consultants to a football team during the summer transfer window, advising teams on which players to buy to improve their roster, hopefully leading to better performance. To achieve this goal, we need to predict what is the price at which a player will be transferred between teams, as this is non trivial and is often different to the market value of the player. Given our predictions for the price of the players, we build a prescription framework in which we recommend decisions to General Managers on which players to keep, sell and acquire. Finally, we run 1000 simulations under different market conditions to simulate competition and run interpretable clustering on the output in order to identify different strategies teams have based on their budget and team conditions. Combined, we present an interpretable end-to-end prescription framework for transfer period action, tailored to the needs of a given team.

2 Data

For the data regarding skills and characteristics of players, we decided to use FIFA as a proxy. This is based on the fact that every year, Electronic Arts (EA) spends significant amounts of money in to making a high quality representation of reality. We have downloaded from kaggle a dataset[1] with player statistics from FIFA 2015 up to FIFA 2022.

We then also need data regarding each player's transfer price. For this, we used a dataset of past transactions of players between teams. We found on Github a repository[2] with the transaction data as recorded by Transfermarkt[3] (which is considered to be the reference website for players transaction information) from the 1992/1993 season to July 2022. It is widely known in the football domain that after the transfer of Gareth Bale from Tottenham Hotspur to Real Madrid in 2013, the transfer market completely changed. For this reason, we were initially planning on using only data from 2013 to train our models, but since the FIFA data with the statistics of the players is only available from 2015, we will only consider data from 2015 on-wards. We created a master dataset joining the statistics coming from the FIFA dataset and transfer price coming from the transfermrkt dataset.

2.1 Data Exploration

In the market there was a generally increasing level of transfer volume in terms of money between teams, until the pandemic hit and this volume significantly reduced. It is clear

[1] https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset

[2] https://github.com/ewenme/transfers

[3] https://www.transfermarkt.us

when looking at the plots of the total volume and average volume of monetary spending across the years in appendix A. It is also generally interesting to explore the relationship between the transfer price and the overall value assigned by FIFA.

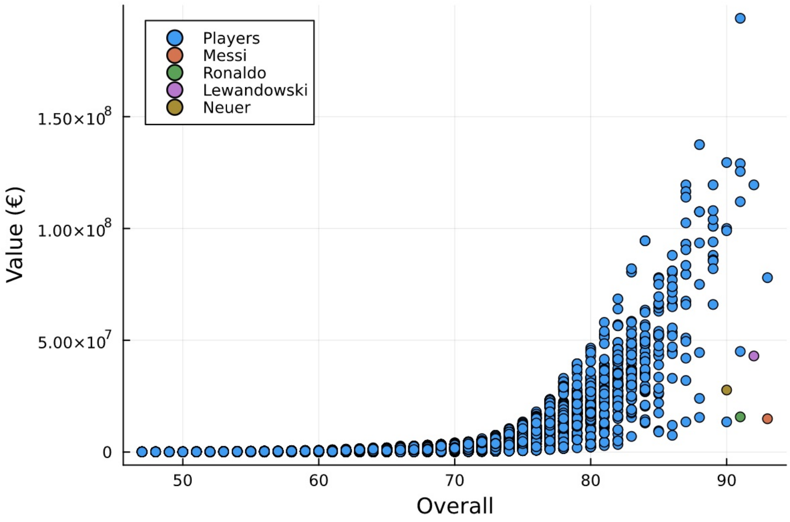

Figure 1: Scatter of Transfer fees and overall value

We can see how the transfer price increases more than linearly with the overall value, with only a few exceptional players that are rated very highly with a relatively low transfer price. This is mainly because these players are very good players but are older, and likely less worth large investment, for example Cristiano Ronaldo, Manuel Neuer, and Robert Lewandowski.

3 Problem Formulation

We will want to maximize a measure that indicates how overall "skillful" a team is, combining the value at the moment and the potential value of the roster. This will be done by creating a metric that measures this combination for every player and then summing over all the players that will compose the team. We want to have the highest possible value subject to some constraints, as, for example, we don't want a team that is too old or too young, we don't want to completely restructure the team, and most importantly, we have to respect the team's budget.

To be able to respect a budget constraint, we would have to know for how much we will buy and/or sell players, and for this we will have to use a Machine Learning framework, able to estimate the price tag of each player. We will then use this prediction to create the budget constraint. Given the uncertainty in the constraint, we add robustness, in order to hedge against worst case scenarios of prediction errors.

Finally, in order to understand the decision making process, we run simulations under various market conditions, and interpret the acquisition decisions made by teams in these simulations using interpretable clustering.

4 Machine Learning Estimation

As for training and validation, we considered all the players that were transferred between 2015 and 2021. The final end-to-end model is tested on players in 2022.

4.1 Modelling

Various methods were applied in order to identify which modelling techniques were most appropriate, as well as to extract meaningful information that can serve General Managers in their decision making. Cross validation was used given the limited training data as well as for hyper-parameter tuning.

4.1.1 Polynomial Regression

Polynomial regression is an extension of linear regression to higher powers. Observing the non-linear relationships in the data, as seen in Figure 2, polynomial regression is a natural method choice. We considered feature transforms of the form x, x^2 and x^3.

4.1.2 Holistic Regression

Given the nature of the data, having more than 100 covariates, it was important to choose carefully what features to use while modelling. For this reason, we implemented Holistic Regression, as one of its key features is that of sparsity. It is further able to identify nonlinear relationships via feature transforms, which is key due to the nonlinear relationships in the data, as seen in Figure 2. For every covariate x, we considered transformations x, x^2, $ln(x)$ and $\sqrt{x}$.

4.1.3 Optimal Regression Trees

In order for General Managers to adopt these methods, interpretability is a key outcome. For this reason, we additionally implement Optimal Regression Trees (ORTs), in order to bring understanding to the key drivers of transfer value. Maximum depth and minimum buckets were tuned as hyper-parameters.

4.1.4 Random Forest

Extending beyond single tree methods, we implement Random Forests for their ability to generalize due to the aggregate of many single tree predictions. Maximum depth, number of trees and minimum buckets were tuned as hyper-parameters.

4.1.5 XGBoost

Finally, due to their success in literature and practice, we implement XGBoost. Maximum depth, number of estimators and minimum buckets were tuned as hyper-parameters.

While not as interpretable as ORTs, Holistic Regression or Polynomial Regression, we can extract feature importance from both XGBoost and Random Forests. Given the importance of interpretability, none of the models selected are "black boxes", in order to retain clarity throughout the decision processes.

5 Optimization Formulation

5.1 Objective function and decision variables

We consider decision variables $\alpha_i \in \{0,1\}$ $\forall i \in [T]$ and $\beta_j \in \{0,1\}$ $\forall j \in [M]$ where T is the size of the team and M is the size of the market. In this setting, α_i is 1 if player i (originally belonging to the team) is chosen retained, and 0 if player i is released. On the other hand, β_j is 1 if the team wants to buy player j (originally not on the team), and 0 if the team does not want to buy player j. Our main metric will be a combination of value and potential value of the players. We want to get players that have an overall good value as a high potential, and we will favor players with a potential that is largely higher than the value. We can create this metric as:

$$f(v, \pi) = (\pi - v) + (\sigma v + (1-\sigma)\pi) \tag{5.1.1}$$

where v is the vector of values, and π is the vector of potential values. The value of σ can be chosen by the General Manager, and depends on their prioritization between the next season and the long term capability of their team. This combination will have to be created both for the players of the team and for the players in the market. For this reason we will differentiate between $v^{(T)}$ and $v^{(M)}$, where $v^{(T)} \in \mathbb{R}^T$ and $v^{(M)} \in \mathbb{R}^M$. The same applies to the vector π[4]. We will then sum over the players belonging to the team and in the market. We will differentiate by position by introducing a binary matrix x, where $x_{ip} = 1$ if player i plays in position p.

We consider a normalization factor Ω_p equal to the number of players that the team has in a given position. This is done to weight down positions in which there usually is a higher number of players (eg. midfielders as opposed to goalkeepers). We can formalize it as:

$$\Omega_p = \sum_{i=1}^{T} x_{ip}^{(T)} \alpha_i + \sum_{j=1}^{M} x_{jp}^{(M)} \beta_j \quad \forall p \in [P] \tag{5.1.2}$$

Given our considerations, we can formulate the objective function as follows:

$$\max_{\alpha,\beta} \quad \sum_{p=1}^{P} \frac{1}{\Omega_p} \left(\sum_{i=1}^{T} x_{ip}^{(T)} \alpha_i f\left(v_i^{(T)}, \pi_i^{(T)}\right) + \sum_{j=1}^{M} x_{jp}^{(M)} \beta_j f\left(v_j^{(M)}, \pi_j^{(M)}\right) \right) \tag{5.1.3}$$

We can clearly see how this formulation, due to the normalization factor Ω_p, is nonlinear. We will work around this problem by looking at the current average number players per position across teams as a proxy.

5.2 Constraints

We now start to consider the constraints of the model. Each team has to respect some budget constraint, namely, they cannot spend more than a budget B in a given market session. Here, the teams will also be able to spend the money that they get from selling current players, namely, this money will add to their budget cap. Given wages w and prices of players ρ, we can formalize this constraint as:

[4] And for every other vector for which it will be meaningful to differentiate

$$\left(\sum_{i=1}^{T}\left(w_i^{(T)}\alpha_i - \left(w_i^{(T)} + \rho_i^{(T)}\right)(1-\alpha_i)\right) + \sum_{j=1}^{M}\left(w_j^{(M)} + \rho_j^{(M)}\right)\beta_j\right) \leq B \tag{5.2.1}$$

Given that the price of the players is a prediction coming from the Machine Learning framework, we will have to add robustness to uncertainty in this framework. We can do so by considering that the ρ vectors will belong to some uncertainty sets:

$$\begin{aligned} \rho^{(T)} &\in U^{(T)} \\ \rho^{(M)} &\in U^{(M)} \end{aligned} \tag{5.2.2}$$

Where the uncertainty sets can be built as:

$$\begin{aligned} U^{(T)} &= \{\rho^{(T)} \in \mathbb{R}^T | \rho^{(T)} = \hat{\rho}^{(T)} - \Delta^{(T)}, \|\Delta^{(T)}\|_\infty \leq \epsilon^{(T)}\} \\ U^{(M)} &= \{\rho^{(M)} \in \mathbb{R}^M | \rho^{(M)} = \hat{\rho}^{(M)} + \Delta^{(M)}, \|\Delta^{(M)}\|_\infty \leq \epsilon^{(M)}\} \end{aligned} \tag{5.2.3}$$

We select this ϵ parameter in order to cover from a certain percentage of error in the validation set.

We can also imagine teams as not wanting to have players that are too old, nor too young, in order to balance experience now and prospects for the future. For this reason, we may want the team to have a mean age between γ (representing the minimum age), and Γ (representing the maximum age). We can write this constraint using an age vector a, formalized as:

$$\gamma \leq \frac{1}{P}\sum_{p=1}^{P}\frac{1}{\Omega_p}\left(\sum_{i=1}^{T} x_{ip}^{(T)}\alpha_i a_i^{(T)} + \sum_{j=1}^{M} x_{jp}^{(M)}\beta_j a_j^{(M)}\right) \leq \Gamma \tag{5.2.4}$$

In general, it is not a good idea to completely revolutionize a team, because new players will likely not have synergies, and they might need time to adapt to new teammates and play styles. For this reason, we can imagine that a team would want to keep a certain percentage Ψ of its team fixed, in order to maintain old synergies between players. This is formalized as:

$$\frac{1}{T}\sum_{i=1}^{T}\alpha_i \geq \Psi \tag{5.2.5}$$

For every role, each team needs a certain number of players, in case of injuries, red cards, or general unavailability to play for certain matches. Given the minimum requirements for every position ψ_p $\forall p \in [P]$, we can write the constraint as:

$$\sum_{i=1}^{T} x_{ip}^{(T)}\alpha_i + \sum_{j=1}^{M} x_{jp}^{(M)}\beta_j \geq \psi_p \quad \forall p \in [P] \tag{5.2.6}$$

We will also bound the number of players that compose the roster to be a number between λ and Λ. We will estimate these values looking at some industry standards and using personal experience. The motivation is simply that teams do not want to have too few players because they would not be able to adapt to different situations during the season, nor too many because it would cause internal conflicts on the minutes played:

$$\lambda \leq \left(\sum_{i=1}^{T} \alpha_i + \sum_{j=1}^{M} \beta_j \right) \leq \Lambda \tag{5.2.7}$$

Lastly, we formalize the integrality of the decision variables:

$$\begin{aligned} \alpha &\in \{0,1\}^T \\ \beta &\in \{0,1\}^M \end{aligned} \tag{5.2.8}$$

In appendix A we present the complete formulation of the optimization problem.

6 Results

In this section, we discuss the performance of the machine learning methods used, as well as present an interpretable prescription framework that suits the constraints and adapts to the needs of different teams.
We can see in the table below the prediction results of the different algorithms used for every category of players.

6.1 Machine learning performance

Models	OSR^2 Att	OSR^2 Mid	OSR^2 Def	OSR^2 Gk
Polynomial Regression	0.363	0.595	0.289	0.624
Holistic Regression	NA	NA	NA	0.270[5]
Optimal Regression Tree	0.478	0.137	0.519	0.528
Random Forest	0.498	0.507	0.602	0.598
XGBoost	0.755	0.607	0.849	0.598

Table 1: Performance table

As we can see from the results, XGBoost dominates the other algorithms on all of the positions, with the exception of the goalkeepers. This is likely due to the smaller training set for goalkeepers, consisting only of 461 players, which does not allow the algorithm to properly learn. Consequently, for this category, we will use the prediction coming from the polynomial regression.

We then took this predictions and plugged them in to the optimization framework.

6.2 Prescription and impact

In a real life, teams are competing for players, meaning that just because you are financially able to sign someone, it does not mean that you will be actually able to sign them, as another team may sign them first. Furthermore, the results of complex optimization is often not interpretable. This makes adoption of analytical methods challenging for high stake decisions in sports. In order to tackle these issues, we run 1000 simulations, each of which the team can only access 10% of the market, achieving two important results:

[5] Holistic regression did not scale for positions with larger datasets. Even for goalkeepers after running for 10 minutes the optimality gap was still 75%

- We simulate competition in the market, developing a proxy for counterfactuals and identifying a list of players valued by the team.
- This creates a target, namely whether or not we buy a player in any simulation. This allows us to employ interpretable clustering, namely, train an Optimal Classification Tree (OCT) to interpret acquisition decisions.

As mentioned, interpretable results are key in order to obtain stakeholder buy-in and implement this framework in practice. To this point, we train an OCT on the results of the above simulation, leveraging interpretable clustering in order to provide stakeholders a glimpse in to the characteristics of players their team is valuing in the transfer market.

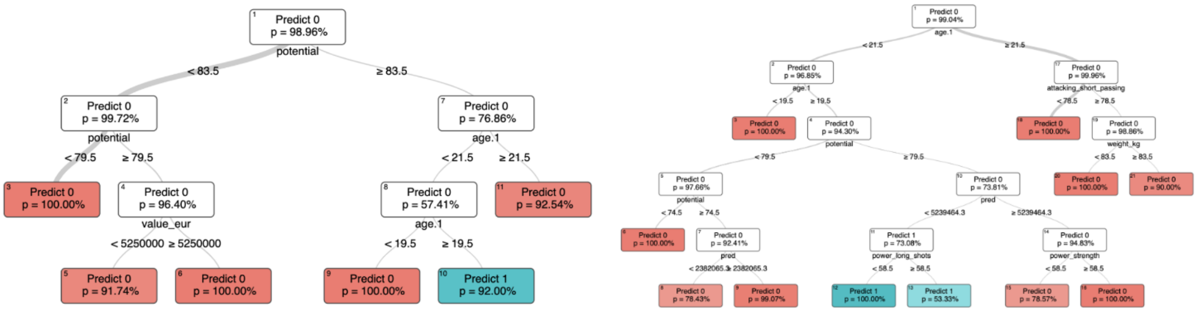

Figure 2: Interpretable clustering on transfer selections made by high budget Barcelona FC (left), and low budget Torino FC (right). Both have accuracy $>$ 99%.

A quick examination of the tree indicates that decision making is similar overall. In this setting (as mentioned, parameters are tuneable by any given team), players aged 19-21 with high potential are clearly the ones each team valued the most. Note however, that the classification for Torino FC is more complex, and actually has lower accuracy. This points to an interesting result, namely, the tighter your budget, the harder your decisions become. With a lower budget, a team may have to settle for possibly lower overall value, and look instead for other good qualities.

We examined two evaluation metrics, beyond the objective value, to compare the players acquired versus those released as a result of our prescription. They are:

- Mean value: $\frac{\sum_p f(v,\pi)_p}{\text{Number of players}}$[6]
- Mean dollar value: $\frac{f(v,\pi)_p}{\sum_p \rho_p}$

Comparing players released versus those obtained[7], rich teams improve their average player value, however worsen their value per dollar by seeking out star players. Conversely, lower budget teams focus more on affordable players with high potential, improving team value per dollar while slightly compromising the overall value[8].

[6] Note here that the subscript "p" indicates the players of formed team, not the position as used before. It for ease of notation.

[7] See appendix C for an example of market session

[8] This was calculated creating the metric for every player and comparing it before and after the market session

It is interesting to note that the market session that we prescribed to the teams is extremely close to the one that they actually performed in terms of releases, and is representative of the interest of the club in terms of acquisitions. Analyzing the market session of Barcelona FC, we can see how the club has actually sold or sent on loan Neto, Piqué, Matheus Pereira, and Braithwaite. Jordi Alba and Sergi Roberto are still with the club, however both are considered as captains of the team, therefore their true value to the team is greater than the value expressed in terms of performance. In terms of acquisitions, the results of the model are not far off in terms of interest as the club expressed interest for the players proposed, but ended up buying different players who possessed similar qualities. This is due to the fact that reality is more complex than what we modeled, as in a real market session there would be a lot of competition between the clubs, all trying to buy the most promising players on the market. This is confirmed observing the prescribed actions to Torino FC, which show similar behaviour. The interaction of competition is then evident if we think about the fact that in real life, Torino FC sold Gleison Bremer, a relatively young, extremely good, high potential player, just because they received an incredibly high offer from Juventus, and thanks to this money they could spend more in the market.

These results show that teams could not only use this model to influence their own decision making, but additionally to gain insight in to how their competition will act, running it using the parameters they believe would be used by competitors.

7 Conclusion

Prescription machine learning is extremely difficult in the context of football transfer action, due to the complexity of generating high quality outcome metrics and counterfactuals. Furthermore, obtaining buy-in from key stakeholders and decision makers relies on both performance and interpretability.

Given more resources and time, adding a dimension of time, prescribing decision making over a multi-year period could be a fruitful. This becomes a very interesting problem as contracts come in to play, and data gains time series elements. Moreover, the extension to include specific positions in the field could be interesting, as well as provide teams more flexibility. Instead of having attackers, midfielders, and defenders, we would expand the framework to be more granular, considering left wing, right wing, central midfielder, and so on[9].

We have developed an interpretable end-to-end pipeline, from data to decisions, leveraging the power of machine learning, robust optimization and interpretable clustering. This data driven prescription method provides football General Managers a tool to help them in their decision making. The flexibility of this framework makes it adaptable, able to fit the needs of any team, regardless of budget.

[9] See appendix D

A Data Analysis

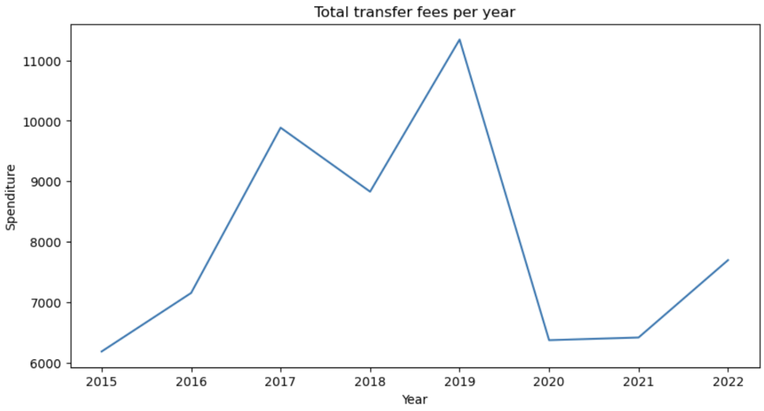

Figure 3: Total transfer fees per year

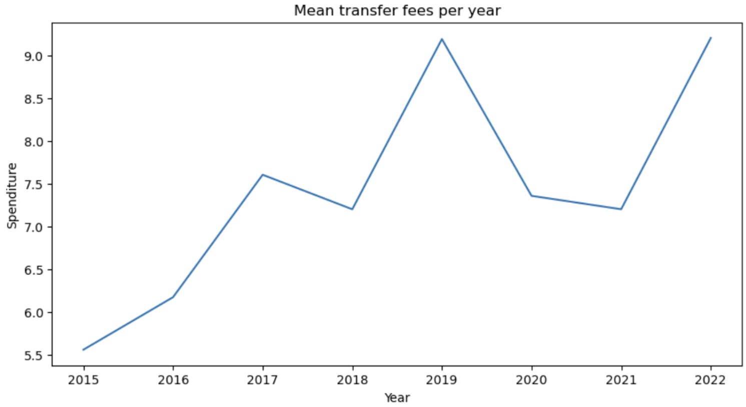

Figure 4: Mean Transfer fees per year

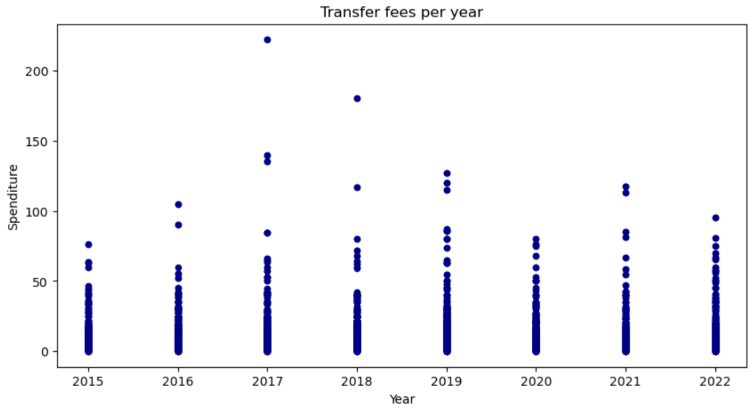

Figure 5: Scatter of Transfer fees per year

B Complete formulation

Combining all of the preceding, we can write the final formulation of the model as follows:

$$
\begin{aligned}
\max_{\alpha,\beta} \quad & \sum_{p=1}^{P} \frac{1}{\Omega_p} \left(\sum_{i=1}^{T} x_{ip}^{(T)} \alpha_i f\left(v_i^{(T)}, \pi_i^{(T)}\right) + \sum_{j=1}^{M} x_{jp}^{(M)} \beta_j f\left(v_j^{(M)}, \pi_j^{(M)}\right) \right) \\
\text{s.t.} \quad & \left(\sum_{i=1}^{T} \left(w_i^{(T)} \alpha_i - \left(w_i^{(T)} + \rho_i^{(T)} \right) (1 - \alpha_i) \right) + \sum_{j=1}^{M} \left(w_j^{(M)} + \rho_j^{(M)} \right) \beta_j \right) \leq B \\
& \gamma \leq \frac{1}{P} \sum_{p=1}^{P} \frac{1}{\Omega_p} \left(\sum_{i=1}^{T} x_{ip}^{(T)} \alpha_i a_i^{(T)} + \sum_{j=1}^{M} x_{jp}^{(M)} \beta_j a_j^{(M)} \right) \leq \Gamma \\
& \frac{1}{T} \sum_{i=1}^{T} \alpha_i \geq \Psi \\
& \sum_{i=1}^{T} x_{ip}^{(T)} \alpha_i + \sum_{j=1}^{M} x_{jp}^{(M)} \beta_j \geq \psi_p \quad \forall p \in [P] \\
& \lambda \leq \left(\sum_{i=1}^{T} \alpha_i + \sum_{j=1}^{M} \beta_j \right) \leq \Lambda \\
& \alpha \in \{0,1\}^T \\
& \beta \in \{0,1\}^M
\end{aligned}
\tag{B.0.1}
$$

C Market Optimization Result

Running the optimization model these are the results that we got:

Acquisitions			
Name	Overall	Potential	Transfer Price
E. Harland	88	93	60.3M
T. Kubo	75	88	10.0M
T. Alexander-Arnold	87	82	56M
G. Donnarumma	89	93	16.4M
Releases			
Name	Overall	Potential	Transfer Price
M. Braithwaite	77	77	6.6M
Matheus Pereira	68	76	5.6M
Jordi Alba	86	86	6.9M
Piqué	84	84	37.7M
Sergi Roberto	81	81	18.6M
S. Umtiti	80	80	6.2M
Neto	82	82	5.0M

Table 2: Prescribed acquisitions and releases to Barcelona FC

And running the optimization for Torino FC these are the results:

Acquisitions			
Name	Overall	Potential	Transfer Price
Afonso Sousa	69	82	480k
A. Perea	65	82	free
Morato	68	84	4.0M
Diogo Costa	73	85	7.5M
Releases			
Name	Overall	Potential	Transfer Price
S. Zaza	76	76	5.6M
S.Verdi	75	75	6.9M
C. Ansaldi	78	78	37.7M
R. Rodríguez	76	76	18.6M
K. Djidji	71	72	6.2M

Table 3: Prescribed acquisitions and releases to Torino FC

D Expanded framework

Here we can see a possible expansion of the attacker, midfielder, defender framework to a more granular definition of the positions in the field. This gives room for a better and more specific decision in the optimization procedure, giving teams the ability of selecting players exactly in the position that they desire, not only generally midfielder, or defender.

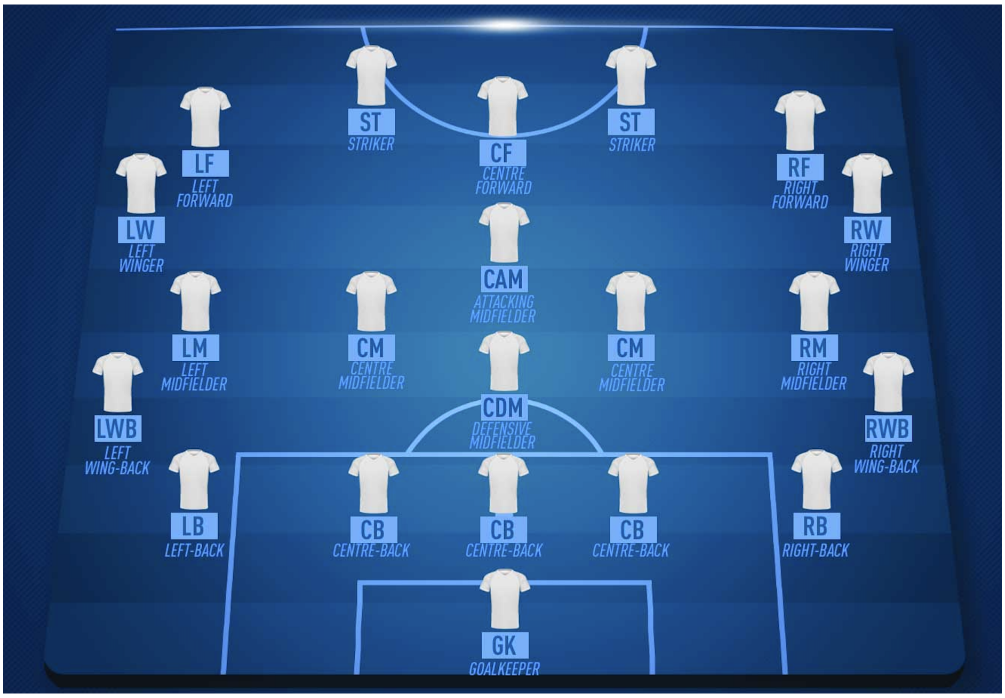

Figure 6: Possible positions in the field

E Team contributions

Both team members contributed equally. While everyone contributed to each stage, many stages were owned by one of the members.

Brainstorming and developing modeling ideas (which models to be used for machine learning prediction task, optimization formulation, simulations and interpretable clustering) were all joint efforts. Additionally all writing and slide creation was shared. Data cleaning and engineering was split evenly as well.

E.1 Marco

As an ex semi-professional player himself, Marco brought to the table the background knowledge required to frame this problem and understand the difficulties that might be faced along the way. He owned data collection as well as coding and implementing the optimization problem.

E.2 Tom

As a terrible soccer player, Tom let Marco focus on the intuition behind decision making. He owned implementing the machine learning algorithms and the simulation / interpretable clustering implementation.