Maximizing Player Market Value and Club Valuation Using Machine Learning: A Strategy for Selling Clubs

Liuxinhao Gao, Akshay Navada, Tanvi Saini, Mike Shu, Nina Wu

Business Understanding

Selling clubs are football clubs that primarily focus on identifying, developing, and nurturing young talent to sell these players to larger, wealthier clubs; some good examples are Benfica, Brighton, and Ajax in Europe. Often, these clubs have strong youth academies and extensive scouting networks. Rather than relying on lucrative sponsorship deals like bigger teams in Europe, the business model of selling clubs revolves around investing in player development and earning revenue through player transfers. To succeed both on the pitch and maintain financial health, these clubs must pinpoint **player skills and attributes that boost market value**. By understanding the specific skills that drive market value, selling clubs can **adapt their development programs** and **tailor their coaching styles** to focus on training players in these high-value areas. This targeted approach can greatly increase the potential transfer revenues and position these clubs as key players in the global football market. This project aims to use supervised machine learning to help selling clubs predict three critical elements:

1. **Which trainable player attributes have the greatest impact on market value?**
2. **Which players are being overvalued so we can cash in while he is still perceived highly by the market**？
3. **How to accurately assess rising stars' value to maximize profit during transfer negotiations with bigger clubs**？

By uncovering these insights, selling clubs can make smarter, **data-driven decisions** about player development, **modifying their training programs** and **coaching methods** to emphasize the skills that add the most market value. This optimization leads to higher market value and。as a result, **higher transfer fees**, better resource allocation, and a more strategic approach to player development. Clubs can shift from a generalized coaching model to a **performance-driven, market-oriented training regimen**, ensuring they maximize the return on investment in their players.

Why This Project is Crucial:

The global football transfer market is a **€6.5 billion industry**, with selling clubs playing a **pivotal role** by supplying big-name teams with top-end talents. For these selling clubs to excel in this rather competitive ecosystem, they must strategically develop players with specific skills that elevate their market value to turn future player transactions into massive profit-making opportunities. This project seeks to provide selling clubs with insights into which player attributes to focus on during training programs, helping them boost financial returns through data-driven insights.

Selling clubs rely heavily on player transfers as a primary source of income, and understanding which player characteristics are most sought after can significantly increase transfer revenues.

- Ajax is a prime example of success, generating a staggering €223 million from player sales in 2022 alone, with a major transfer being Frenkie de Jong to FC Barcelona for €75 million (after acquiring him for just €1 million), yielding an astonishing 7500% ROI. By better understanding the skills that made De Jong so valuable, Ajax could replicate this success with future talents.

Data Understanding

Our project is based on the dataset “Football Players’ Transfer Fee Prediction Dataset” on Kaggle: https://www.kaggle.com/datasets/khanghunhnguyntrng/football-players-transfer-fee-prediction-dataset

This dataset contains 2,061 players’ data from 17/18 season to 20/21 season across 134 different performance metrics besides the information such as “Name”, “Value”, and “Position”. Each column in the dataset is a player-season combination — for instance, “Passes Attempted (18/19)”.

Key metrics such as Goals, Assists, and Minutes Played indicate offensive contribution, playmaking ability, and player reliability. More advanced stats like Completed Progressive Passes and Completed Crosses into the Penalty Box offer insights into a player's ability to contribute to team strategy and break through defenses. From a business perspective, these metrics enable selling clubs to identify high-potential players, optimize development strategies, and maximize transfer value by focusing on attributes that drive market demand.

One significant bias in the dataset is omitted variable bias, especially the absence of goalkeeping stats like saves, save percentage, and clean sheets. This gap makes it challenging to identify key attributes that determine a goalkeeper’s value, therefore, we decided against developing models for goalkeepers. Additionally, other omitted variables such as height and weight might also influence a player's market value but are neglected.

Data Preparation

We cleaned our data in the following ways to better run our desired machine-learning models:

Missing Values: We have many missing values (NAs) for players for certain columns across different seasons. These missing values need to be treated before moving forward with the processing of data.

- For NA values that can be explained with logic - for instance, if a player made 0 passes and therefore 0 successful passes, his passing completion% would have a denominator of 0, resulting in NA calculations - we replaced them with 0
- For players who have an entire season's stats missing (either due to being out of contract or being a youngster), we decided to take a weighted average approach.

***Wide data-set weighted average approach*:** Our dataset contains player statistics from four seasons, spanning 548 columns. Since player value typically reflects performance across multiple seasons, we transformed the dataset into a weighted format for efficient modeling.

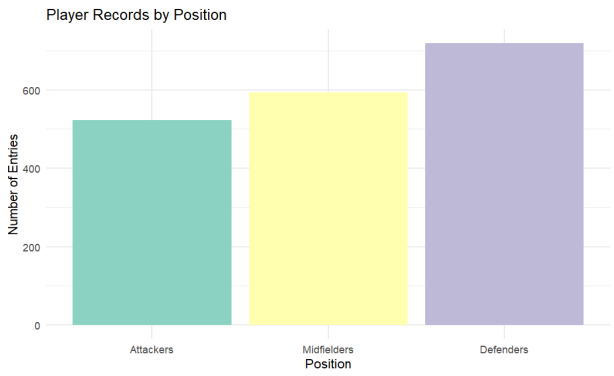
- We compiled various player statistics across four seasons into one using a weighted average that prioritizes recent data: 50% for the most recent season (20/21), 25% for 19/20, 15% for 18/19, and 10% for 17/18. This method also accounts for players with missing seasonal data, as the total weight would be adjusted accordingly.

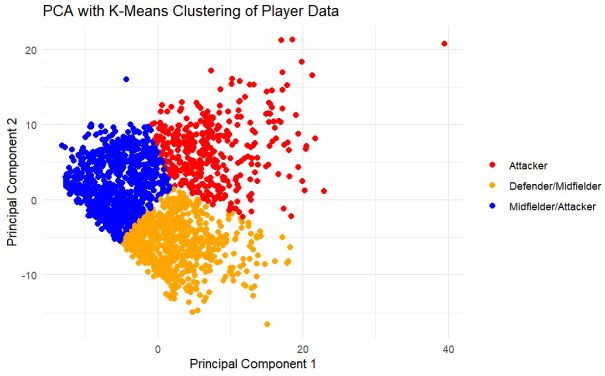
Column Names: We reconstructed some variable names by replacing brackets with hyphens to avoid any syntax or runtime errors.

Appropriate Data types: We ensured that each column was stored in appropriate data type by converting string/character columns such as "position", and "contract years left" into dummy variables to better suit our machine learning models.

Player Duplicates: We identified and removed duplicate player entries, retaining only the most recent based on the player's age and remaining contract duration.

Sub-datasets: We anticipate different clusters based on player positions, as factors that go into evaluations of attackers are likely to be different from those of defenders. We performed a combination of PCA, k-mean, and elbow method, resulting in 3 clusters which is likely to do with player positions.





To address this issue, we created four sub-datasets on player position (attack, midfield, defenders) so that each model is tailored to specific attributes and performance metrics relevant to each position.

Modeling

We decided to run three machine learning algorithms - **Linear Regression**, **Lasso Regression**, and **Random Forest -** with "Value" being our dependent variable on three of our sub-datasets. We decided to use k-fold with 10 folds to systematically test the robustness of each model.

Multiple linear regression is straightforward and allows testing direct relationships between player attributes and value, but multicollinearity can be an issue with a large number of variables in our dataset. Lasso reduces model complexity by penalizing less important attributes, making it less prone to multicollinearity, but it can miss some features if the penalty is too high. Random

Forest handles overfitting well and provides insights into attribute importance through its ensemble of decision trees, but is less interpretable compared to the other models.

By understanding which attributes most significantly impact a player's market value based on the coefficients in our model, selling clubs can tailor their training programs to enhance these key skills. This focus on high-impact areas can lead to increased player performance that the market values and, consequently, a higher market valuation.

With prediction models that the three machine learning algorithms come up with, these clubs can also time their sales more effectively. If a player's market value is way above his predicted value, clubs should consider capitalizing on this moment to secure the best possible transfer fees. Vise versa, this strategy also reduces the risk of underpricing talent.

Evaluation

We decided to evaluate the performances of our models using Out-Of-Sample (OOS) Root Mean Squared Error (RMSE), which measures the extent to which predicted values differ from actual values by calculating the square root of the average squared differences. Lower RMSE values indicate a better model fit. After selecting the best-performing model, we will be able to identify player attributes with the highest coefficients, which are the most influential attributes in driving player value.

We found that for attackers, Lasso Regression provided the best prediction model with an RMSE of **10638169** (lower compared to the linear model's 15165764 and rf's 11881900). In fact, Lasso performed the best for midfielders and defenders as well.

Trainable attributes for Attackers:

Metric	Value
xG/90_Weighted_Avg	4746817.8166
xGandxA/90_Weighted_Avg	4361853.3317
Non-PenaltyGoals_Weighted_Avg	821964.8579
Gls_Weighted_Avg	365048.1532
PenaltyKicksWon_Weighted_Avg	307959.2912
RedCards_Weighted_Avg	300023.2651
Non-PenaltyxGandxA/90_Weighted_Avg	245634.5094
CarriesintoAttackingPenaltyBox_Weighted_Avg	231470.4638
TouchesinAttackingPenaltyBox_Weighted_Avg	19405.0714
AerialDuelLost_Weighted_Avg	-7771.5092

Coaching Strategy for Attackers

For attackers, lasso suggests that the most impactful attributes with the largest coefficients include the likes of expected goals per 90 minutes (xG/90), non-penalty goals, and penalty kicks won. Therefore, one key coaching tactic for attackers should be to focus on improving attackers' proficiency at scoring. This can be achieved by working on finishing drills, one-on-one attacking scenarios, and quick passing combinations in the final third to create more goal-scoring opportunities, as expected goals and assists per 90 minutes is also an important attribute determined by the model. Additionally, focusing on carries into the penalty box and improving their ability to receive passes successfully under pressure will further enhance their goal-scoring potential and increase their value in the market.

Trainable attributes for Midfielders:

Metric	Value
xA/90_Weighted_Avg	16219976.9551
Gls_Weighted_Avg	579714.0180
Ast_Weighted_Avg	557739.8789
PenaltiesConceded_Weighted_Avg	1345268.0213
FoulsDrawnLeadingtoGoals_Weighted_Avg	547911.6464
TotalAssists_Weighted_Avg	499201.0969
PassesLeadingtoGoals_Weighted_Avg	441224.4999
ShotsonTarget/90_Weighted_Avg	369000.9080
xG_Weighted_Avg	349414.4869

Coaching Strategy for Midfielders

Based on lasso, midfielders who significantly contribute to creating goal-scoring opportunities tend to have higher values. Key attributes such as expected assists per 90 minutes (xA/90) and passes leading to goals, which represent chance creation skills, stood out specifically. Therefore, selling clubs should focus on training their midfielders with drills that revolve around goal-creating actions. By developing these skills, selling clubs can enhance the performance and value of their midfielders, leading to higher transfer fees. In addition to targeted training, one interesting observation from our model is that the Premier League dummy variable has a huge positive impact on midfielders' values (4.389559e+06), which corroborates with previous key attributes and suggests that more attacking midfielders in a league with more end-to-end actions are valued higher.

Although Lasso had the lowest OOS RMSE for defenders, its variables and coefficients are not

very interpretable. Therefore, we went with the random forest model to analyze trainable attributes, with “Overall” being the aggregate measure of an attribute’s importance across all trees, which indicates how much it contributes to the prediction accuracy.

Trainable attributes for Defenders:

Metric	Overall
NumberofTimesReceivedPass_Weighted_Avg	100.000000
NumberofTimesPlayerwasPassTarget_Weighted_Avg	96.040040
PassesCompleted-Allpass-types_Weighted_Avg	36.084686
TotalCarries_Weighted_Avg	26.372652
TotalDistanceCarriedtheBall_Weighted_Avg	22.661878
TouchesinMidfield3rd_Weighted_Avg	11.207714
ShotsLeadingtoGoals_Weighted_Avg	11.111466
%AerialDuelsWon_Weighted_Avg	10.120521
PassCompletion%-Allpass-types_Weighted_Avg	8.960492

Coaching Strategy for Defender

For defenders, the rf model highlights that the most important attributes that count towards a player’s value include the number of times received passes, total carries, and touches in the midfield 3rd. This is quite a surprise find! This implies that coaches should put more emphasis on enhancing defenders’ positioning and technical decision-making, which is greatly in line with the trend of ball-playing defenders that we see currently. This can be achieved through drills that emphasize reading the game and transition practices from defense to offense. Specific drills such as small-sided games that simulate real match scenarios and passing drills under pressure can also improve defenders’ ability to receive and distribute the ball effectively.

Aside from training, using our data-driven model, a selling club like Brighton can more accurately gauge its players' market values. For instance, Leandro Trossard, an attacker who plays for Brighton, has a value of €15,300,000, but our model predicted that he is worth €19,394,226. With this insight, Brighton can set a more appropriate asking price if bigger clubs come knocking. By leveraging data-driven predictions, Brighton can confidently negotiate higher transfer fees, ensuring they maximize their RIO.

Deployment

With our prediction models and insights into significant attributes that impact player value, selling clubs can leverage this analysis to better understand and enhance player marketability. By implementing these models, clubs can tailor training programs to focus on attributes that increase market value, ensuring players develop skills that are most valuable in the transfer market, assess accurately players' value, and make informed decisions when it comes to player sales.

One glaring drawback to the model is the lack of appropriate player attributes such as height, weight, and all goalkeeping statistics. It would be beneficial for selling clubs to obtain more adequate datasets going forward.

Focusing too much on specific metrics can pressure players to prioritize certain skills over their overall development as footballers, which could lead to neglection of broader skills. As the game of football evolves and styles of play change, relying solely on specialized skill sets may not necessarily be a sustainable future-proof strategy. One way of mitigating it for the selling clubs is to ensure a balanced training approach that prioritizes player health, well-being, and holistic development.

Appendix

Contribution of each team members:

Liuxinhao Gao: data selection, report and slides refinement

Akshay Navada: data selection, report write-up, slides creation, RStudio modeling

Tanvi Saini: data selection and cleaning, report and slides refinement

Mike Shu: data selection and cleaning, report write-up, slides creation, RStudio modeling and visualization

Nina Wu: data selection, report and slides refinement