FROM PITCH TO PRICE

A MACHINE LEARNING APPROACH TO FOOTBALL PLAYER VALUATION

1. **Introduction**

The aim of this project is to build a predictive model for **estimating the market value of professional football players** based on a dataset containing personal, technical, and financial attributes. The following analysis includes data preprocessing, exploratory data analysis, feature selection, and supervised learning.

Firstly, we focused on managing redundant variables through correlation analysis. Then, we placed emphasis on performance, ensuring reproducibility, and applying cross-validated model evaluation to enhance generalization. Redundant or causally ambiguous variables, such as the "release clause", are explicitly excluded to maintain temporal and structural validity, as it is better explained later on.

The modeling phase involves the comparison of different algorithms, trained and validated through 5-fold cross-validation to mitigate overfitting, with Gradient Boosting selected as the final predictive model.

This report presents the rationale behind each step, from variable selection to model assessment.

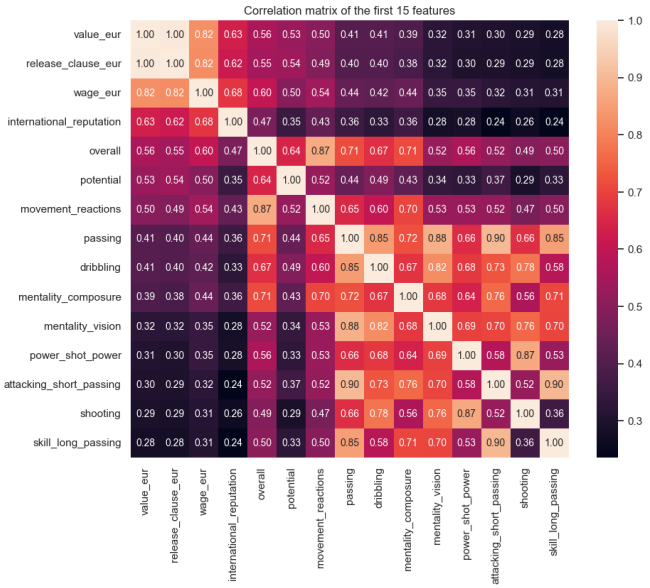
2. **Feature Selection**

2.1 **Correlation Matrix**

To assess the linear relationships between features and the target variable (value_eur), a Pearson correlation matrix was computed. Among the continuous features, *overall, wage_eur,* and *release_clause_eur* showed the **highest positive correlation** with the target. Based on redundancy observed, we retained only a reduced set of informative features to avoid multicollinearity and dimensionality inflation. This analysis revealed several pairs of variables exhibiting near-linear relationships. These are feature pairs with very high mutual correlation (typically $|r| > 0.9$), indicating that they encode largely redundant information.

In particular, technical attributes such as *dribbling, short_passing,* and *ball_control* were all strongly correlated with the global rating feature *overall*. Including all such variables would introduce multicollinearity, impairing model stability and interpretability, especially for linear methods. To address this, **we retained only the most informative feature from each highly correlated group** (*overall*), and discarded the twins. This approach reduced dimensionality, mitigated noise, and preserved the predictive power of the selected subset.

We proceeded to make scatter plots between *value_eur* and the most correlated variables chosen from the previous point, taking account for twin explanatory variables, which can be found in the notebook



Here we visualized the top 20 correlated features with the target using a heatmap to better interpret the structure of inter-feature relationships.

2.2 **Dealing with Release clause**

Although the variable *release_clause_eur* exhibits an extremely strong correlation with the target variable *value_eur*, **it was intentionally excluded from the predictive model**. Despite its apparent predictive utility, we observed that **model performance remained comparable even in its absence**. More importantly, the decision to exclude it was motivated by real-world relevance and causal coherence.

In fact, in real football operations, a player's market value is first assessed based on objective attributes such as performance metrics, position, age, and physical characteristics. The release clause is **then** defined by the club after this valuation step, often as a strategic or contractual decision rather than a direct reflection of the player's intrinsic worth. Moreover, it is a pure estimate, deriving from the valuation given to the current value and the expectations about the future, therefore being possibly not truly reflective of the real value. Driven by market dynamics, negotiation leverage, or deterrent purposes, these clauses have become increasingly inflated in recent years, rendering them less representative of actual value.

Moreover, we wanted the model to suit different needs. Assume there is a manager renewing a contract, who does not have access to an up-to-date release clause. To determine some financials of the contract, including the release clause, he will probably need to assess the monetary value of the player. Not using the release clause in evaluating this value makes the model more flexible for real-life applications.

Including such a variable would not only introduce temporal inconsistency and reflect post commercial decisions that can vary widely and arbitrarily across clubs and leagues. In fact, it would introduce a variable whose determination may be a consequence of the target variable. **To maintain the model's generalizability and real-world applicability, we prioritized features grounded in player characteristics** rather than administrative factors.

2.3 **Wage**

Similar reasoning to the one applied with the release clause has been applied to the *wage*. Although wage_eur had a high correlation with the target variable *value_eur*, we chose not to include it in our model. **In real-world scenarios, a player's wage is usually the *result* of their estimated value**. Wages are negotiated

based on many external factors (club finances, agent demands, contract length) which makes them unreliable as a direct indicator of value. **To keep our model realistic and useful for practical applications, we focused only on features that are available and meaningful before a player's value is determined.**

2.4 **Dummy Variables for player_position and league_name**

In the dataset, the *player_positions* variable lists all the roles a footballer can occupy on the field. Since each player may appear in multiple positions, we simplified this information by **selecting only the primary position**, defined as the first-listed role. While there are around fifteen standard positions in football, **we grouped them into ten broader tactical categories to reduce dimensionality and enhance interpretability**. These groups were defined as follows: central and forward attackers (ST/CF/LF/RF), wingers (LW/RW), central attacking midfielders (CAM), wide midfielders (LM/RM), central midfielders (CM), defensive midfielders (CDM), wing backs (LWB/RWB), full backs (LB/RB), centre backs (CB), and goalkeepers (GK). **The grouped positions were then encoded using one-hot encoding, resulting in ten dummy variables**. This transformation allowed us to capture essential tactical information about the player while ensuring compatibility with both linear and non-linear models and avoiding the ambiguity associated with multi-role players.

In parallel, we introduced a binary variable to account for the competitive level of the league in which each player competed. The original ***league_name*** **variable was converted into a dummy feature** to identify players active in the five most prestigious European leagues, as defined by FIFA rankings. Specifically, players in the Spanish Primera División, English Premier League, Italian Serie A, German Bundesliga, or French Ligue 1 were assigned a value of 1, while those in all other leagues received a 0. **This variable captures the reputational and financial impact of league affiliation on market value** without introducing unnecessary categorical complexity.

2.5 Determining the number of years until the contract ends

The dataset refers to the early 2022 football season. This can be understood by the fact that the variable club_joined showed that the latest recorded club entry occurred in 2021. To compute the number of years remaining on a player's contract, we subtracted the current year (2021) from the *club_contract_valid_until* variable. This calculation resulted in the creation of a new variable: *years_to_end_contract*.

2.6 **Handling of Missing Data**

Some of the rows presented missing values for the *value_eur* feature due to them not having a club. However, the number of players presenting a missing *value_eur* was small enough to motivate our decision to discard them, without significantly compromising the model's performance.

Finally, we applied a threshold-based criterion whereby any variable with more than 15% missing values would be excluded from the dataset. However, none of the selected features exceeded this threshold, so all were retained for further analysis. If one of the selected variables had spurious missing data, the missing values were **filled using the median of that feature.**

2.7 **Distribution of Key Features**

We analyzed the distribution of the three key monetary variables in the dataset *value_eur, wage_eur*, and *release_clause_eur,* **both in their original and log-transformed forms**. All three variables initially exhibit **extreme right skewness,** with skewness values exceeding 6, indicating that most observations are concentrated at low values, with a long tail extending to the right.

To address this, **we applied a log transformation**, which significantly **reduced skewness**: *value_eur* dropped from 7.80 to 0.57, *wage_eur* from 6.45 to 0.47, and *release_clause_eur* from 8.03 to 0.62. These **transformations helped to normalize the distributions, making them more symmetric and closer to Gaussian.**

While log transformation is not strictly necessary for the Gradient Boosting model eventually selected, since it handles non-normal distributions well it remains a useful step for improving the interpretability of the data, stabilizing variance, and reducing the influence of extreme outliers.

2.8 **Five-fold cross validation**

To ensure model robustness and reduce the risk of overtuning, all candidate algorithms were evaluated using 5-fold cross-validation. In this procedure, the dataset was partitioned into five equally sized folds: in each iteration, four folds were used for training and one for validation. This cycle was repeated five times, allowing training data to serve as validation.

This provided a **more stable estimate of model performance, averaging out variability** due to data partitioning. It also helped to **prevent overtuning** by ensuring that model tuning was not biased by a single train-test split. Finally, enabled fair comparison among models, supporting an informed selection of the best-performing algorithm.

3. **Model Selection**

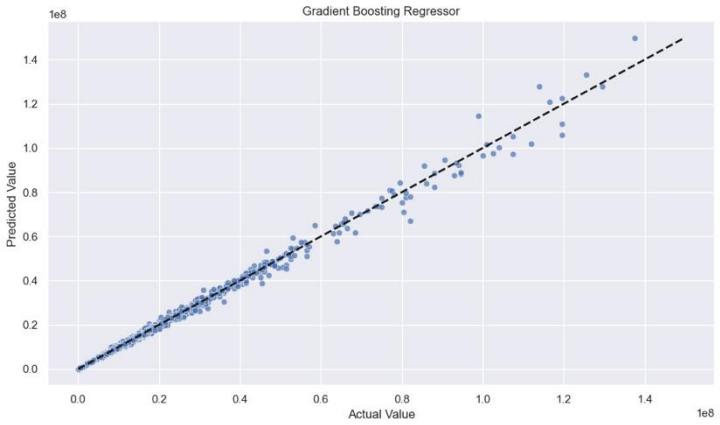
We selected **Gradient Boosting Regressor (GBR) as our final model** due to its ability to capture non-linear relationships and interactions between features. The initial model was trained using a curated subset of predictors, including player attributes (*overall, potential, age, international_reputation, years_to_end_contract*), dummy variables for the grouped player positions, and the binary indicator for top-league affiliation. The first baseline configuration used *n_estimators*=100, *learning_rate*=0.1, and *max_depth*=3. Performance on the test set was evaluated using Root Mean Squared Error (RMSE) and R^2 score, and visually validated with a scatter plot comparing predicted vs. actual values.

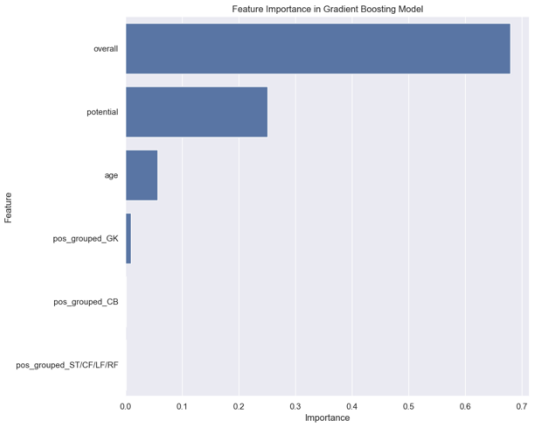
To assess model stability and avoid overfitting, we then applied 5-fold cross-validation, calculating the mean and standard deviation of RMSE and R^2 across folds, to prevent overtuning, as doing a simple 80/20 split led to very inconsistent results, due to the relatively small amount of the data and the large amount of outliers. This procedure was repeated with a second configuration using *n_estimators*=400, *learning_rate*=0.2, and *max_depth*=3, resulting in improved average performance.

We then **conducted a fine-tuned hyperparameter optimization using GridSearchCV**, testing combinations of *n_estimators* (400–600), *learning_rate* (0.12–0.20), subsample (0.70–0.90), and *min_samples_split* (3–10), with *max_depth* fixed at 3. The **best configuration was selected based on 5-fold cross-validated RMSE**. The optimized model was then evaluated on the test set, yielding improved metrics over previous configurations. To ensure robustness, we further validated this final model using cross-validation, confirming consistent performance across folds.

The best performance recorded by our model on the test data coming from the train/test split, after finetuning, is represented by an RMSE of about 450 thousand, and an R^2 score of 0.9964, with a random state set to 42.

Note that, to ensure reproducibility should one want to exactly reproduce some peculiar results, every main function that has been employed accepts an optional parameter for setting a specific randomization state.





4. **Feature Importance Analysis**

The **feature importance analysis** from our gradient boosting model reveals that a player's overall rating is by far the most influential factor in determining market value, accounting for approximately 65% of the model's predictive power. Player potential ranks second with about 25% importance, indicating that future development prospects significantly impact valuation. Age contributes roughly 5% to the model's decisions, while position-specific variables have minimal individual impact. This aligns with real football industry valuation practices, as current ability (overall rating) and growth potential are primary considerations, with age affecting how much development runway remains for a player.