# Football Player Price Predictor

Samsudeen Afolabi & Elisa Cerdá Doñate

# Background and Objective

In the dynamic and competitive landscape of professional football, accurately assessing the market value of players is a complex challenge. **Traditional methods often rely on subjective evaluations**, and with the increasing influence of data in sports, there's a growing need for a more objective and data-driven approach to predicting football player values. This problem becomes even more pronounced as clubs seek to optimize their investments, negotiate contracts, and strategically build winning teams.

## Objective

**The primary objective of this project is to develop a model for predicting the market value of football players based on relevant attributes and performance metrics.** By leveraging advanced statistical and machine learning techniques, we aim to provide clubs, agents, and stakeholders in the football ecosystem with a reliable tool that enhances decision-making processes related to player acquisitions, contract negotiations, and overall team-building strategies.

*Data Source : https://www.kaggle.com/datasets/mubarak2000/european-football-players-market-value*

# Problem Approach

**Our approach involves analyzing comprehensive dataset containing a wide range of player attributes, performance statistics, and market values**. Through feature engineering, statistical analysis, and machine learning algorithms, we aim to identify the key factors influencing a player's market value and build a predictive model that can generalize well to new data.

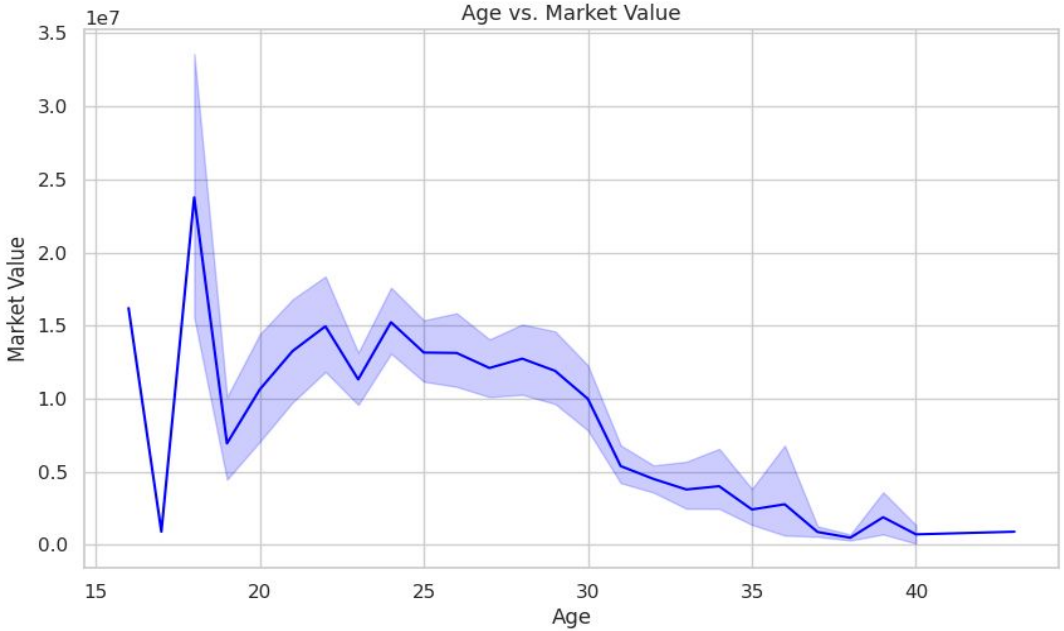
## Key components of Approach

- **Exploratory Data Analysis**
- **Data Cleaning**
- **Data Visualization**
- **Data Treatment**
- **Model Building**
- **Model Evaluation**

# Player Value vs Amount of Goals per Player

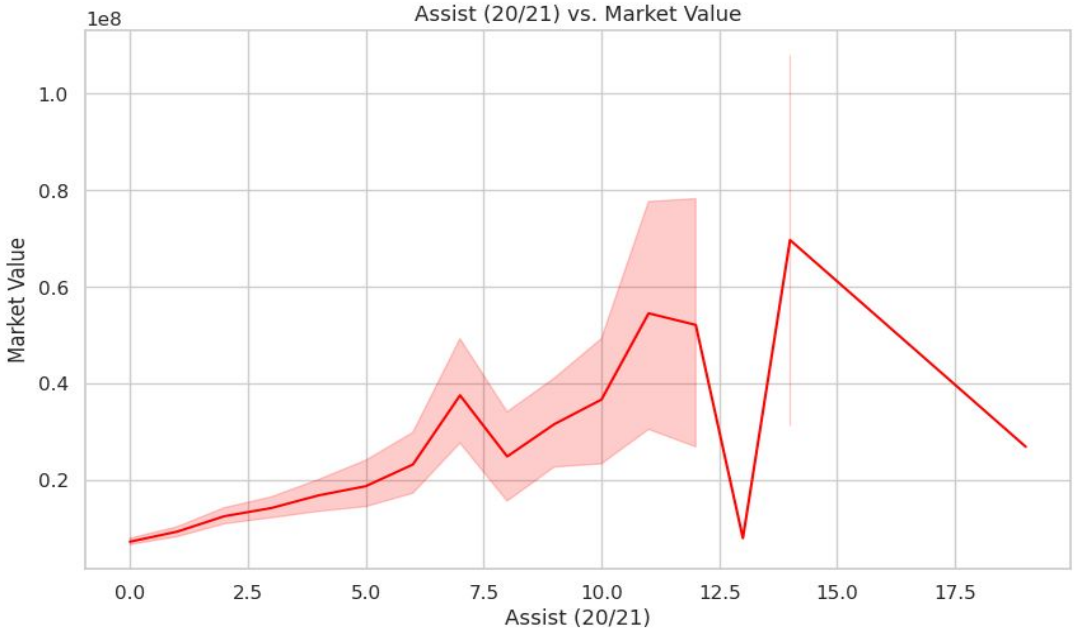

- A trend is observed where the value of players increases with the number of goals
- However, a deviation of this trend occurs from 27 goals onwards
- This anomaly could be attributed to the scarcity of players achieving such high goal counts

# Player Value vs Player Age



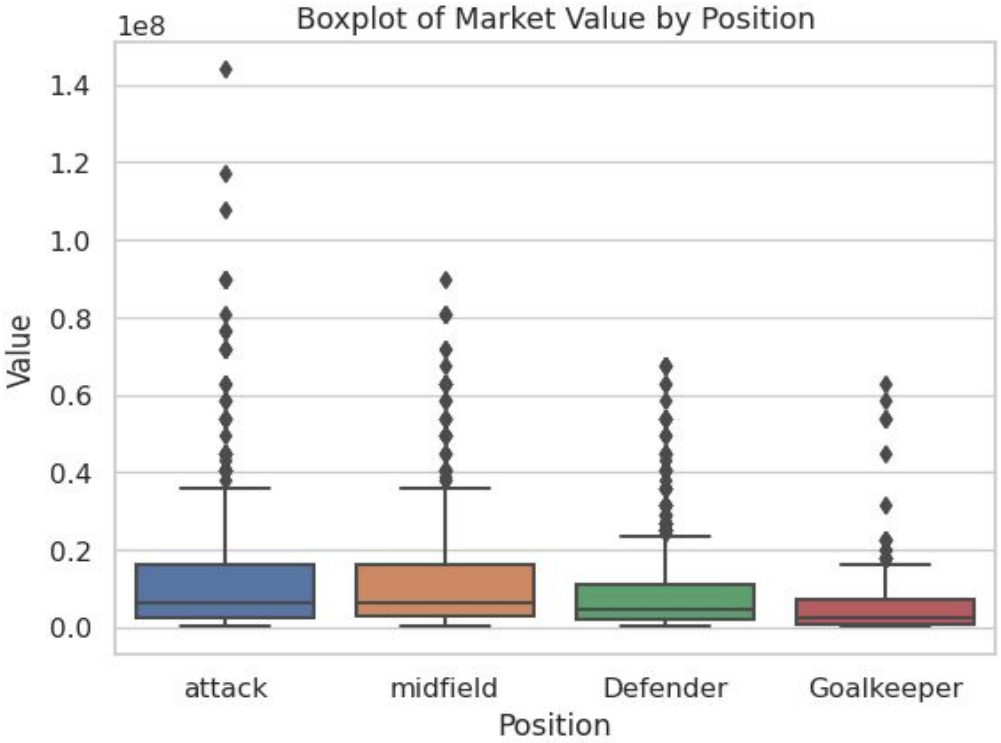

- Distinct pattern in player valuation based on age
- High value for 17-22y:
  team's anticipation on their careers
- High value for 22-30y:
  prime years in their career
- Stark decline beyond 30y

# Player Value vs Amount of Assists per Player




- Mostly positive linear relationship between player value and assist amount
- Player with higher amount of assists tend to have higher values
- An anomaly to this trend is observed beyond 13 as not so many players have very high assist numbers

# Player Value vs Player Position

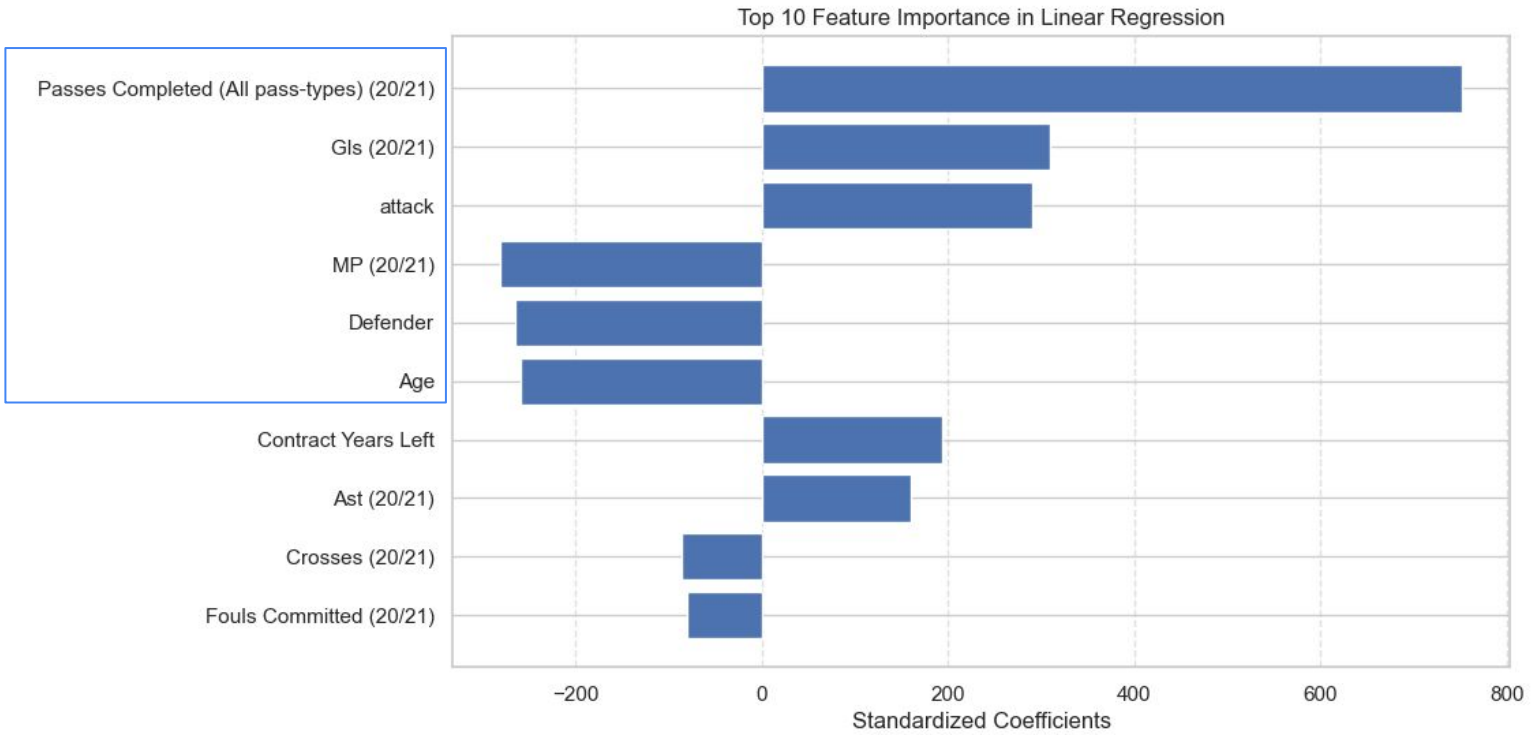



- Cleat trend between player value and player position
- On average, attackers command higher market values compared to midfielders, defenders, and goalkeepers
- This insight highlights the premium placed on attacking positions in the valuation of soccer players.

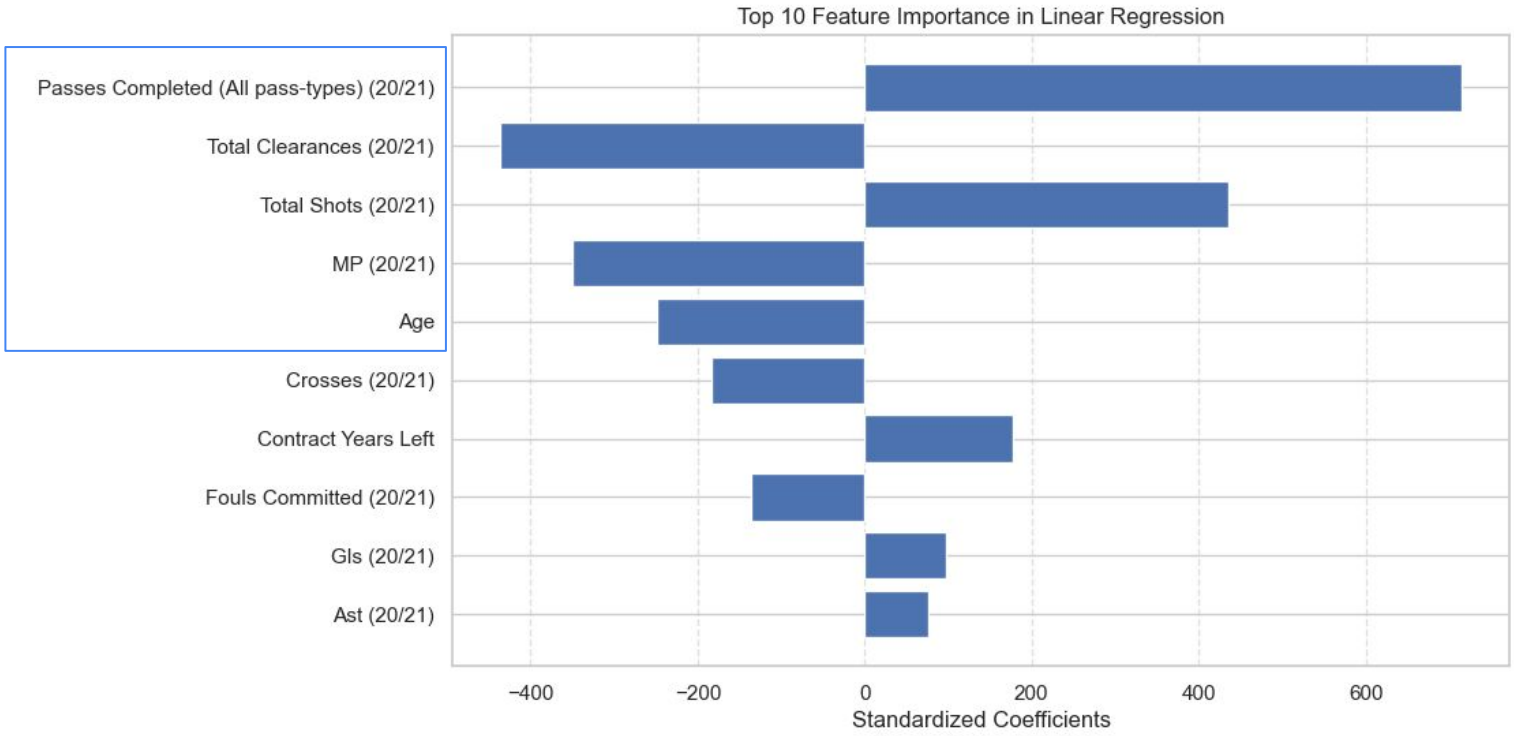
# Data Pre-Processing for Model Building

- **Application of Independent Numerical Variable Threshold** — 0.7/0.8/0.9 correlation threshold
- **Target Variable Re-Scaling** — “Value” divided by 10,000
- **Numerical Independent Variable Scaling and Transformation** — StandardScaler on “Age”; PowerTransformer on remaining variables
- **Nominal Independent Variable Hot Encoding** — Hot encoding on “Position”
- **Training-Testing Data Splitting** — Data partition based on a 75-25 split

# Feature Importance - 0.7 Threshold



# Feature Importance - 0.9 Threshold



# Feature Importance for Player Value Calculation

*Higher variable weight on player value*

*Lower variable weight on player value*

| Model - 0.7 Threshold | Model - 0.9 Threshold |
|---|---|
| **Passes Completed** | **Passes Completed** |
| **Goals** | ***Total Clearances*** |
| **Attack** | **Total Shots** |
| ***Matches Played*** | ***Matches Played*** |
| ***Defender*** | ***Age*** |
| ***Age*** | |
| $R^2 = 0.40$ | $R^2 = 0.41$ |

The $R^2$ of 0.40-0.41 indicates that the model accounts for 40-41% of the variance in player values

# Conclusion

- **“Passes Completed” appears to be a key variable affecting player value**
- **Correlation astringency affects the variable weight, but results on similar model accuracy**
- **Current model fails to accurately predict target variable**
- **The studied variables are not sufficient to fully explain player value**
  (other variables such as actual club and country performance will probably have a high impact on player value)
- **More player data is needed, other variables should be studied or other modelling strategy should be followed**

# Football Player Price Predictor

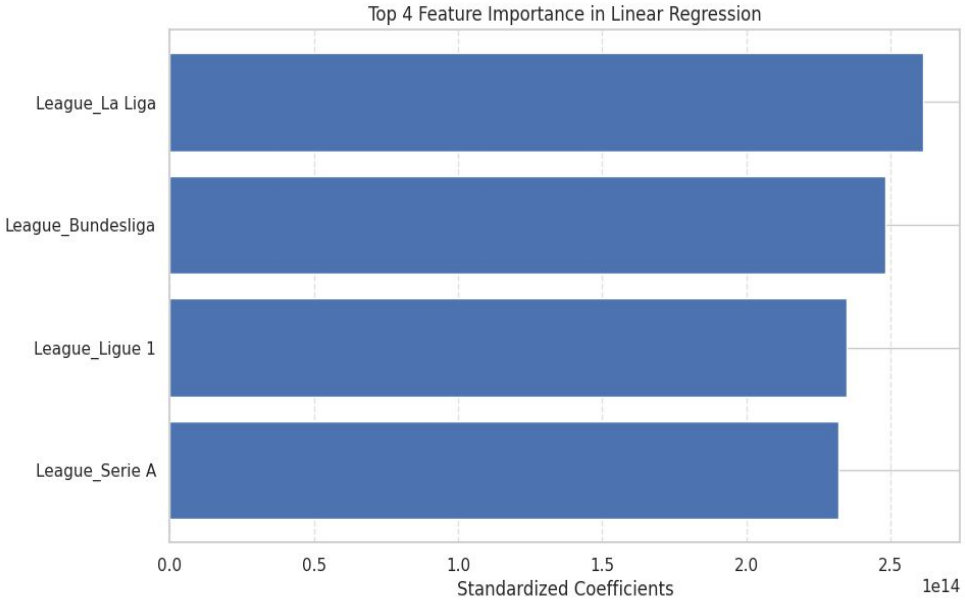
Samsudeen Afolabi & Elisa Cerdá Doñate

# Modelling and Evaluation Metrics

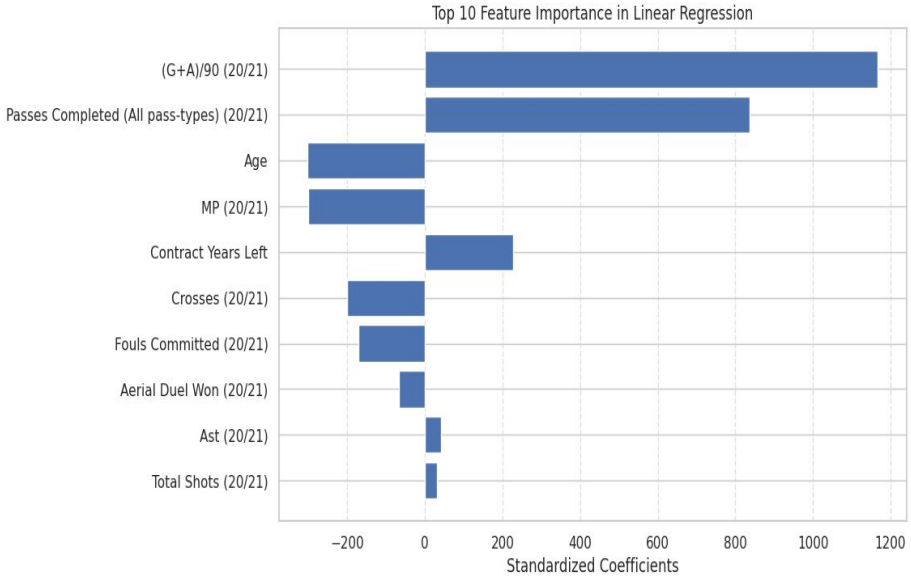
| | Model | |
|---|---|---|
| | 0.7 Threshold | 0.9 Threshold |
| MSE | 550,238 | 545,024 |
| RMSE | 742 | 738 |
| MAE | 561 | 561 |
| $R^2$-Score | 0.40 | 0.41 |

The $R^2$ of 0.40-0.41 indicates that the model accounts for 40-41% of the variance in player values

# Features Importance




*Feature importance for the 0.7 threshold*




*Feature importance for the 0.9 threshold*

# Feature Importance

Top 10 Features Importance Explanation on the 0.9 threshold model:

1. G/A per 90 (Goals/Assists per 90 minutes):
   - ➢ Positive Standard Coefficient: Indicates a positive correlation with the target variable, "Value."
   - ➢ Interpretation: As the G/A per 90 value increases, the player's market value tends to increase. This suggests that players contributing more goals and assists per 90 minutes are valued higher in the market.

2. Total Shots:
   - ➢ Positive Standard Coefficient: Suggests a positive relationship with player value.
   - ➢ Interpretation: Players who take more total shots are associated with higher market values, potentially reflecting goal-scoring ability or offensive contribution.

3. Passes Completed (All pass-types):
   - ➢ Positive Standard Coefficient: Indicates a positive impact on player value.
   - ➢ Interpretation: Players with a higher number of completed passes across different types contribute significantly to team play, positively influencing their market value.

4. Contract Years Left:
   - ➢ Positive Standard Coefficient: Positive correlation with player value.
   - ➢ Interpretation: The number of years remaining on a player's contract positively affects their market value, suggesting that longer contractual commitments contribute to higher valuations.

5. Assists:
   - ➢ Positive Standard Coefficient: Suggests a positive association with player value.
   - ➢ Interpretation: Players providing more assists demonstrate playmaking skills, positively impacting their market value.

# Feature Importance *continued*

1. Total Shots on Target:
   - ➢ Positive Standard Coefficient: Positive relationship with player value.
   - ➢ Interpretation: Players who consistently hit the target with their shots may be perceived as more accurate and valuable, influencing their market worth.

2. Age:
   - ➢ Negative Standard Coefficient: Indicates a negative correlation with player value.
   - ➢ Interpretation: Older players may experience a decrease in market value, reflecting the common trend where younger players are often valued more for their potential and longevity.

3. Matches Played (MP):
   - ➢ Negative Standard Coefficient: Suggests a negative impact on player value.
   - ➢ Interpretation: Higher match participation might not necessarily contribute positively to market value, possibly indicating that overplayed players could be less valued.

4. Crosses:
   - ➢ Negative Standard Coefficient: Suggests a negative relationship with player value.
   - ➢ Interpretation: The number of crosses made by a player might not significantly influence market value or could be perceived as less valuable compared to other contributing factors.

5. Fouls Committed:
   - ➢ Negative Standard Coefficient: Indicates a negative correlation with player value.
   - ➢ Interpretation: Players committing fewer fouls may be perceived as more disciplined or valuable, positively impacting their market worth.

# Modelling and Evaluation Metrics

Linear Regression Modeling:

Model Choice: Applied Linear Regression as the predictive modeling algorithm.

Linear Regression is suitable for predicting numeric values and can provide insights into the relationships between variables.

Evaluation Metrics

0.7 Threshold Metrics:

- Mean Squared Error (MSE): 5.38e+23
- The MSE measures the average squared difference between predicted and actual values. The large MSE suggests significant variance in predictions, indicating the model struggles to capture the variability in the target variable.
- Root Mean Squared Error (RMSE): 7.33e+11
- The RMSE is the square root of the MSE, providing a more interpretable metric in the original units. The extremely high RMSE suggests considerable errors in predicting player values.
- Mean Absolute Error (MAE): 4.74e+10
- The MAE represents the average absolute difference between predicted and actual values. The substantial MAE indicates a lack of accuracy in predicting player values.
- R-squared (R2): -3.31e+17
- The R2 score measures the proportion of variance in the dependent variable explained by the model. The negative R2 suggests a poor fit, and the model does not perform better than a simple mean.

0.9 Threshold Metrics:

- Mean Squared Error (MSE): 5.38e+23
- Root Mean Squared Error (RMSE): 7.33e+11
- Mean Absolute Error (MAE): 667.37
- R-squared (R2): 0.46

Interpretation:

- The 0.9 threshold exhibits substantially improved metrics compared to the 0.7 threshold.
- MAE: The MAE of 667.37 suggests, on average, predictions deviate by approximately 667.37 from the true values. This is a much more reasonable error compared to the 0.7 threshold.
- R2: The R2 of 0.46 indicates that the model accounts for 46% of the variance in player values. While not extremely high, it suggests a moderate level of explanatory power.

Model Interpretability:

- The 0.9 threshold model performs better in terms of accuracy and explanatory power.
- The positive R2 implies that the model, at this threshold, can explain a significant portion of the variability in player values.
- It's crucial to interpret predictions cautiously, considering the inherent complexity of factors influencing player valuations in the football industry.

# Approach

Our approach involves analyzing comprehensive dataset containing a wide range of player attributes, performance statistics, and market values. Through feature engineering, statistical analysis, and machine learning algorithms, we aim to identify the key factors influencing a player's market value and build a predictive model that can generalize well to new data.

## Key components of Approach

- Data cleaning
- Exploratory Data Analysis
- Data Curation & Transformation
- Correlation Matrix
- Scaling
- Data Splitting
- Feature Importance
- Modelling
- Evaluation Metrics

# Problem Statement


In the dynamic and competitive landscape of professional football, accurately assessing the market value of players is a complex challenge. Traditional methods often rely on subjective evaluations, and with the increasing influence of data in sports, there's a growing need for a more objective and data-driven approach to predicting football player values. This problem becomes even more pronounced as clubs seek to optimize their investments, negotiate contracts, and strategically build winning teams.

# Objective

The primary objective of this project is to develop a model for predicting the market value of football players based on relevant attributes and performance metrics. By leveraging advanced statistical and machine learning techniques, we aim to provide clubs, agents, and stakeholders in the football ecosystem with a reliable tool that enhances decision-making processes related to player acquisitions, contract negotiations, and overall team-building strategies.

*Data Source : https://www.kaggle.com/datasets/mubarak2000/european-football-players-market-value*

# Data Cleaning


- Consistent Data-Types for String Columns
- Domain Knowledge and Research-Informed Decisions
- Column Removal

# Data curation and Correlation matrix

1. Target Variable Transformation:
   - ➢ Recognizing the need for scale adjustment, the target variable "Value" was divided by 10,000, effectively making it exponential 4.
   - ➢ This transformation facilitates more manageable numerical values, potentially improving the model's interpretability and convergence.
2. Correlation Matrix Examination:
   - ➢ Conducted a comprehensive correlation analysis to understand the relationships between different variables, especially concerning the target variable "Value."
   - ➢ Detected instances where certain independent variables exhibited strong correlations with each other.
3. Correlation Thresholding:
   - ➢ Established a threshold of 0.7 as a criterion for identifying highly correlated variables.
   - ➢ Implemented a systematic approach to drop one of the columns involved in correlations of 0.7 or higher, reducing multicollinearity effects.
   - ➢ Similarly, applied a stricter threshold of 0.9 to identify and remove variables with exceptionally high correlations.
4. Model Performance Evaluation:
   - ➢ Fed both versions of the dataset, one with a correlation threshold of 0.7 and another with 0.9, into the model.
   - ➢ Evaluated the model performance for each dataset to understand the impact of correlation-based feature removal on predictive accuracy.

# Scaling and Data Splitting

1. Handling Skewed Data:
   - ➢ Power Transform: Applied power transformation to address skewness in the data.
   - ➢ This technique helps normalize the distribution of skewed features, enhancing the performance of linear regression models that assume normality.
2. Nominal Column Encoding:
   - ➢ One-Hot Encoding: Processed nominal columns through one-hot encoding.
   - ➢ Converted categorical variables into a numerical format, creating binary columns for each category.
   - ➢ Facilitates the inclusion of categorical information in the linear regression model.
3. Feature Scaling:
   - ➢ Standard Scaling for Age: Utilized standard scaling specifically for the "Age" variable.
   - ➢ Standardization ensures that "Age" values are transformed to have a mean of 0 and a standard deviation of 1.
   - ➢ This is crucial for linear regression, as it mitigates the impact of variable scales on the model coefficients.
4. Data Splitting:
   - ➢ Train-Test Split: Partitioned the dataset into training and testing sets using a 75-25 split.
   - ➢ Allocating 75% of the data to the training set allows the model to learn patterns from a substantial portion of the dataset.
   - ➢ The remaining 25% serves as the test set, enabling unbiased evaluation of the model's generalization performance.