Soccer Player Valuation

Data Science for Sports

By:

Abhijit Nelli

Jash Patel

Karthik Iyer

Mohit Jain

December 7, 2024

FALL 2024

Table of Contents

Soccer - The Billion-Dollar Game

Football is not just a sport but a multi-billion-dollar industry where clubs invest vast sums to acquire top talent. The transfer market regularly sees record-breaking deals, with clubs fiercely competing for the best players. In such a competitive environment, accurately determining a player's market value is essential.

The Problem

Traditional player valuations are influenced by subjective assessments and often fail to account for all performance aspects. This creates challenges such as:

- **Overvaluation** of players based on reputation rather than performance.
- **Undervaluation** of talented players who lack popularity.
- **Inconsistent metrics** that vary across leagues and analysts.

Stakeholders:

- **Who?** Clubs, agents, analysts, and fans are eager to evaluate player worth.
- **What?** Valuation blends skills, performance, marketability, and potential.
- **Why?** Smarter investments mean competitive advantage.

Key Questions

- **What drives a player’s market value?**
- **How can clubs make smarter, data-driven valuation decisions?**

Our Solution

We propose a data-driven approach using advanced soccer analytics frameworks:

1. **SPADL (Soccer Player Action Description Language)**: Standardizes soccer event data.
2. **VAEP (Valuing Actions by Estimating Probabilities)**: Quantifies the impact of player actions on game outcomes.

We use these frameworks to calculate **Offensive Box Plus-Minus (OBPM)** and **Defensive Box Plus-Minus (DBPM)**. A **Random Forest Regressor** then predicts player market values based on these metrics and **age**.

Data Source

Our dataset is derived from **Kaggle's FIFA Dataset,** which contains:

- **20+ Sub-Attributes** (e.g., Finishing, Sprint Speed, Interceptions).
- **6 Main Attributes**:
    1. **Pace**
    2. **Shooting**
    3. **Passing**
    4. **Dribbling**
    5. **Defending**
    6. **Player Physic**

Additional details include:

- **Player Age**
- **Club Affiliation**
- **Position**

	player_id	fifa_version	short_name	long_name	player_positions	overall	potential	value_eur	wage_eur	pace	shooting	passing	dribbling	defending	physic	age	height_cm	weight_kg	league_id	league_name	league_level	club_team_id	club_name	national
0	158023	15	L. Messi	Lionel Andrés Messi Cuccittini	CF	93	95	100500000.0	550000.0	93.0	89.0	86.0	96.0	27.0	63.0	27	169	67	53.0	La Liga	1.0	241.0	FC Barcelona	
1	20801	15	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	LW, LM	92	92	79000000.0	375000.0	93.0	93.0	81.0	91.0	32.0	79.0	29	185	80	53.0	La Liga	1.0	243.0	Real Madrid CF	
2	9014	15	A. Robben	Arjen Robben	RM, LM, RW	90	90	54500000.0	275000.0	93.0	86.0	83.0	92.0	32.0	64.0	30	180	80	19.0	Bundesliga	1.0	21.0	FC Bayern München	
3	41236	15	Z. Ibrahimović	Zlatan Ibrahimović	ST	90	90	52500000.0	275000.0	76.0	91.0	81.0	86.0	34.0	86.0	32	195	95	16.0	Ligue 1	1.0	73.0	Paris Saint-Germain	
4	167495	15	M. Neuer	Manuel Peter Neuer	GK	90	90	63500000.0	300000.0	NaN	NaN	NaN	NaN	NaN	NaN	28	193	92	19.0	Bundesliga	1.0	21.0	FC Bayern München	

Figure 1: Understanding Our Dataset

Data Preparation

1. Data Cleaning

- Addressed missing values.
- Removed redundancies and standardized formats.
- Ensured consistent currency conversion for market values.

2. Feature Engineering

- **OBPM Calculation**: Aggregated Pace, Shooting, Passing, and Dribbling.
- **DBPM Calculation**: Aggregated Defending, Physic, and Pace.
- **Position-Specific Weighting**: Tailored weights for different positions (e.g., higher Shooting weight for Strikers).

3. Normalization

- Ensured attribute weights were consistent across different positions.

player_positions	count
CM	3303
CB	3272
ST	2979
CDM	2244
LM	2204
RM	2095
CAM	2041
RB	2016
LB	1869
GK	1775
RW	911
LW	839
CF	667
LWB	100
RWB	82

Figure 2: Player Position Distribution

Exploratory Analysis & Data Visualization

Attribute Correlation Analysis

Heat Maps:

- **General Correlation Matrix**: Shows relationships between the 6 main attributes.
- **Strikers**: Strong correlation between Shooting and Dribbling.
- **Defenders**: High correlation between Defending and Physic.

General Correlation matrix –

- It examines the relationships across all player attributes, identifying universally impactful traits for performance and market value.

Key Insights:

- Strong correlations between *potential* and *value (EUR)* highlight that players with higher potential are often valued more.
- Performance metrics like shooting, dribbling, and passing have moderate positive correlations with offensive contributions (*OBPM*), which directly impacts value.
- Defensive attributes such as tackling and interceptions correlate strongly with *DBPM*, showcasing their importance for defensive roles.

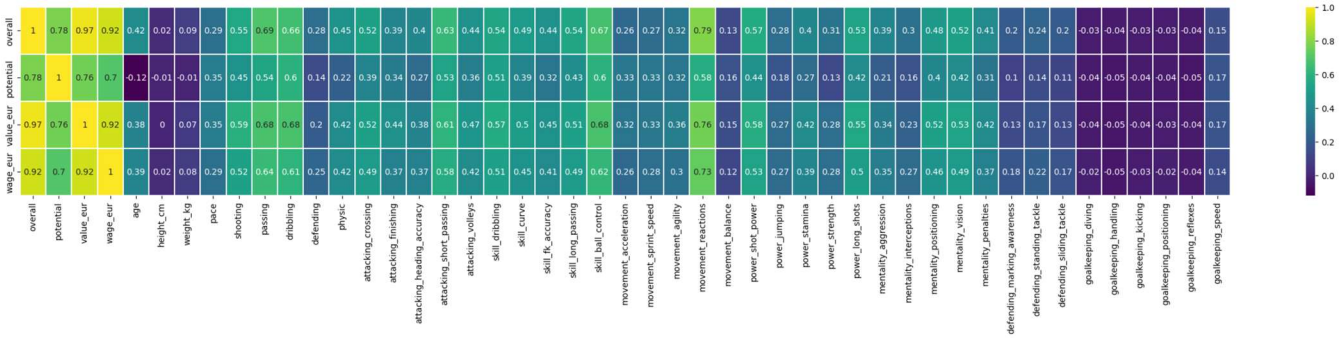


Figure 3: General Correlation Matrix

Goalkeeper Correlation Matrix –

- Focuses on attributes critical for goalkeepers, isolating metrics like diving, reflexes, and positioning.

Key Insights:

- High correlation between *reflexes* and *diving* indicates that agile goalkeepers often excel in stopping shots.
- Moderate correlations between *positioning* and both *reflexes* and *diving* suggest that strategic placement complements physical skills.
- Goalkeeper-specific metrics show weak or no correlation with offensive attributes (e.g., dribbling or shooting), underlining the specialized nature of this position.

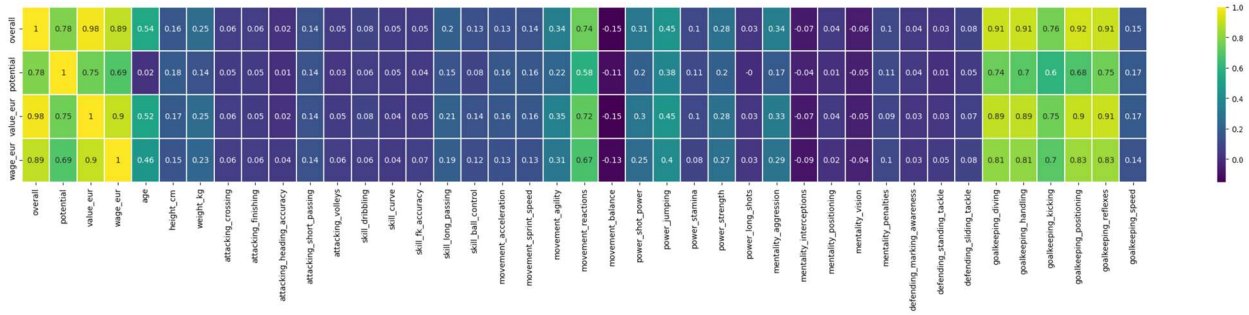


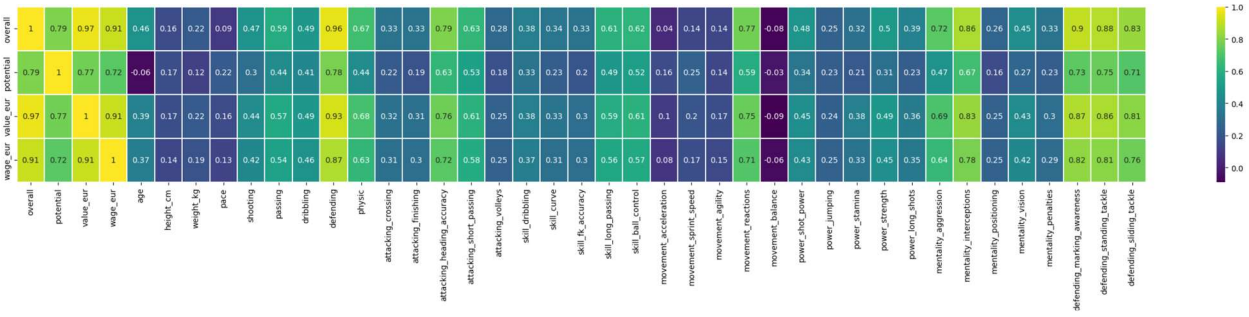
Figure 4: Goalkeeper Correlation Matrix

Center Back (CB) Correlation Matrix –

- Evaluates attributes relevant to defensive roles like strength, standing tackles, and interceptions.

Key Insights:

- Strong correlations among *strength*, *standing tackles*, and *interceptions* indicate these attributes work together to define defensive effectiveness.
- Moderate correlations with *DBPM* highlight the role of these metrics in influencing defensive impact.
- Weak correlations with *OBPM* emphasize the limited offensive role of center-backs.



Striker (ST) Correlation Matrix

- Concentrates on offensive metrics like finishing, positioning, and sprint speed to understand scoring potential.

Key Insights:

- High correlations between *finishing* and *positioning* suggest that scoring efficiency depends on strategic placement as much as technical skill.
- Moderate correlations between *sprint speed* and *dribbling* highlight the role of mobility in creating goal-scoring opportunities.
- Strong positive correlation with *OBPM* reflects how offensive contributions drive a striker's impact and, ultimately, their value.

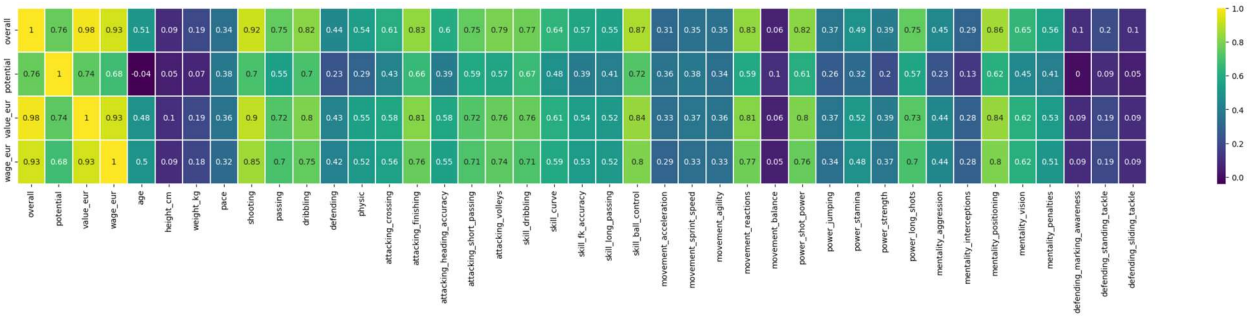


Figure 6: Goalkeeper Correlation Matrix

Applications of the Correlation Matrices:

- **Attribute Selection:** Helps in identifying key variables for predictive models, like *OBPM*, *DBPM*, and *potential*, which strongly correlate with market value.
- **Position-Specific Analysis:** Allows a tailored evaluation of players based on their roles, isolating attributes that matter most for different positions.
- **Strategic Insights:** Reveals how certain attributes (e.g., *finishing* for strikers or *interceptions* for defenders) impact player valuation and performance.

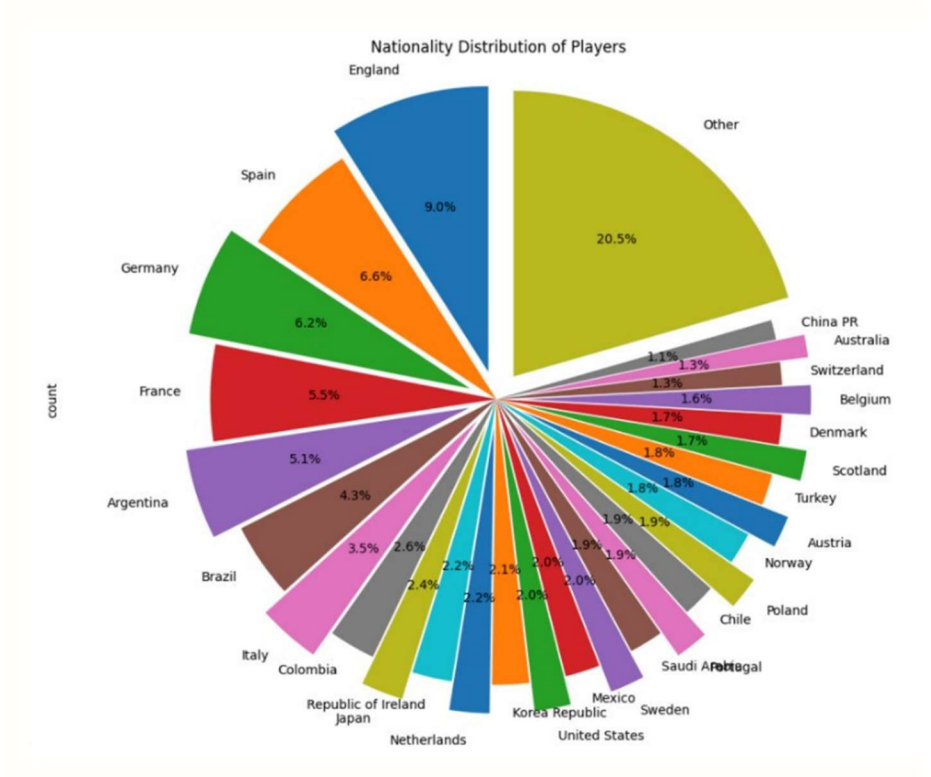


Figure 7: Player Distribution Nation-wise

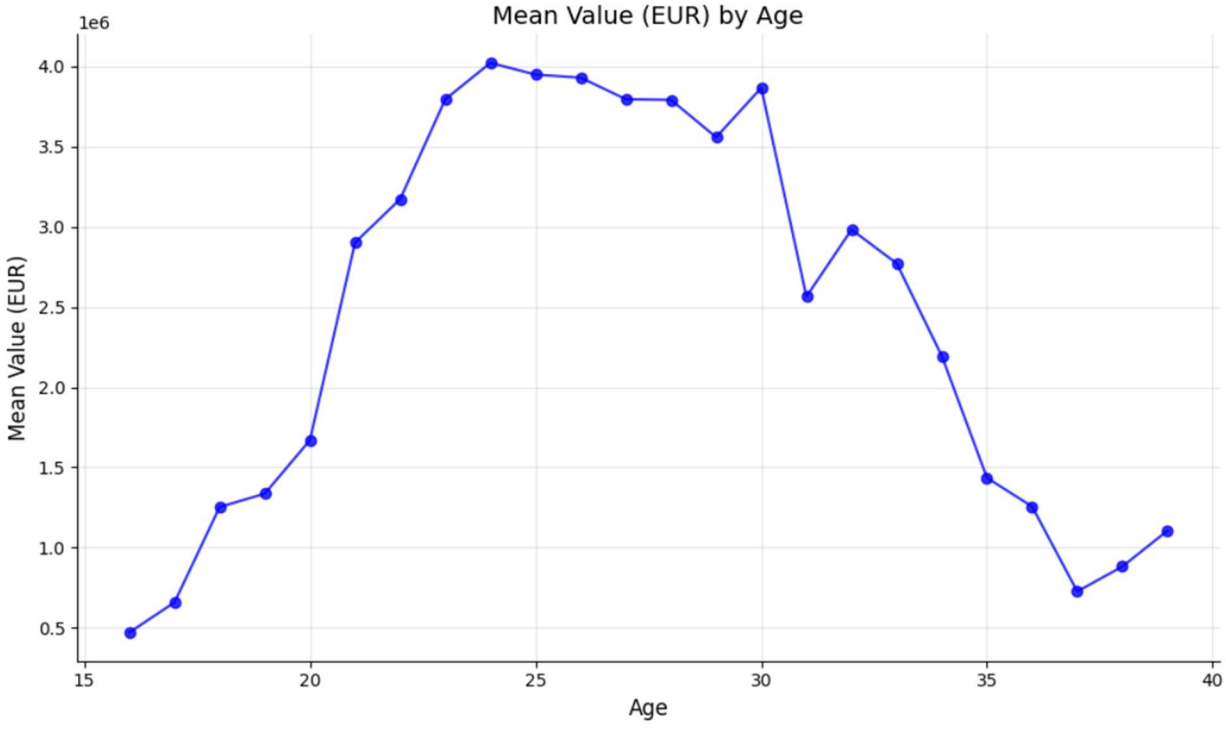


Figure 8: Age vs Player Value - Shows market value trends peaking between ages **24-28.**

Player Value Across Different Positions

Insight

Different positions exhibit varying market values due to the unique skill sets required. Forwards generally have higher market values due to their goal-scoring potential, while goalkeepers and defenders often have lower valuations.

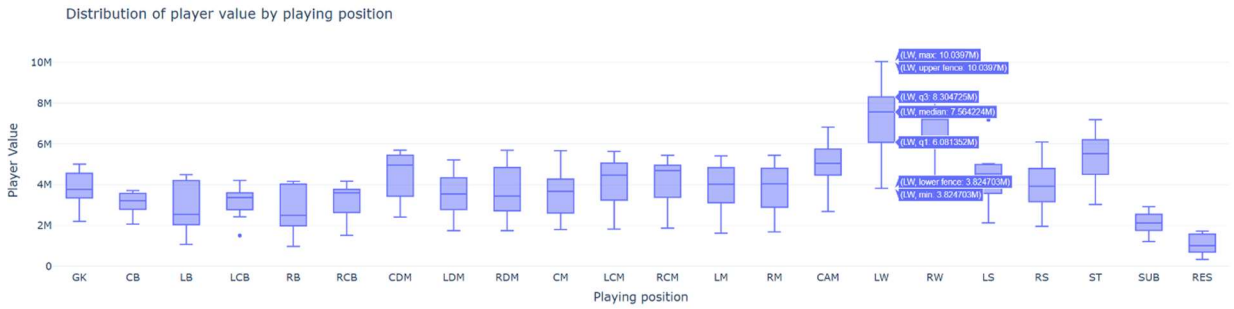


Figure 9: Average Player Value by Position

Analysis:

- **Strikers (ST)** and **Wingers (RW/LW)** tend to have the highest average market values.
- **Centre-Backs (CB)** and **Goalkeepers (GK)** have comparatively lower market values.

Player Value Across Different Nations

Insight

Nationality can influence market value due to factors like player popularity, league reputation, and talent pipelines.

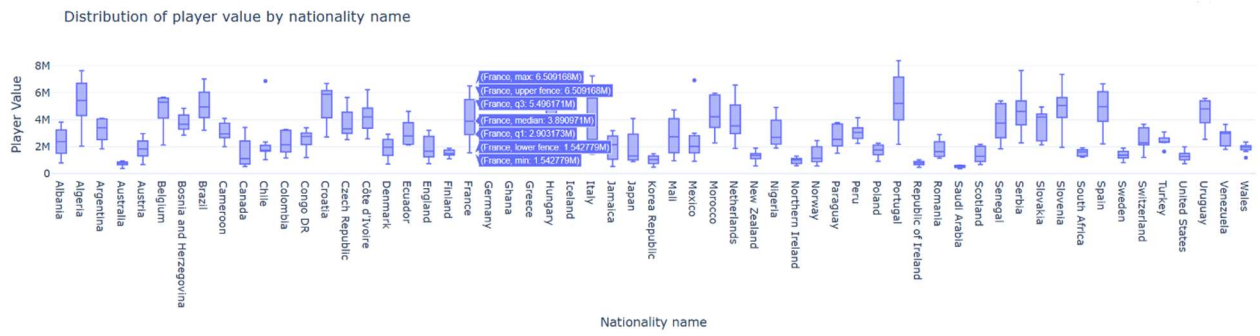


Figure 10: Average Player Value by Nation

Model Selection: Random Forest Regressor

We selected a **Random Forest Regressor** for its ability to:

1. Handle non-linear relationships.
2. Provide feature importance analysis.
3. Perform well with structured data.

OBPM and DBPM Analysis

OBPM and DBPM calculation –

```
play_attributes = ['pace', 'shooting', 'passing', 'dribbling', 'defending', 'physic']

obpm_weights = {}
dbpm_weights = {}
obpm_norm = {}
dbpm_norm = {}

for position in data_cleaned['player_positions'].unique():

    position_data = data_cleaned[data_cleaned['player_positions'] == position]

    corr_matrix = position_data[play_attributes + ['value_eur']].corr()

    value_corr = corr_matrix['value_eur'].drop('value_eur')

    obpm_attributes = ['pace', 'shooting', 'dribbling', 'passing']
    dbpm_attributes = ['defending', 'physic', 'pace']

    obpm_weight_raw = value_corr[obpm_attributes]
    dbpm_weight_raw = value_corr[dbpm_attributes]

    obpm_weight_normalized = obpm_weight_raw / obpm_weight_raw.sum()
    dbpm_weight_normalized = dbpm_weight_raw / dbpm_weight_raw.sum()

    obpm_weights[position] = obpm_weight_raw.round(2)
    dbpm_weights[position] = dbpm_weight_raw.round(2)

    obpm_norm[position] = obpm_weight_normalized.round(2)
    dbpm_norm[position] = dbpm_weight_normalized.round(2)

OBPM = pd.DataFrame(obpm_weights)
DBPM = pd.DataFrame(dbpm_weights)

OBPM_N = pd.DataFrame(obpm_norm)
DBPM_N = pd.DataFrame(dbpm_norm)
```

```python
import numpy as np
import pandas as pd

# List of player attributes
o_play_attributes = ['pace', 'shooting', 'passing', 'dribbling']
d_play_attributes = ['defending', 'physic', 'pace']

# Initialize dictionaries to store the OBPM and DBPM calculations
obpm_values = []
dbpm_values = []

# Loop through each player
for index, row in data_2022.iterrows():
    # Get the player position (could be multiple positions, so we handle that)
    positions = row['player_positions'].split(', ')
    player_obpm = 0
    player_dbpm = 0

    # Loop over positions for each player
    for position in positions:
        # Get the OBPM and DBPM weights for the position from the DataFrames (OBPM_N and DBPM_N)
        try:
            position_obpm_weights = OBPM_N.loc[:, position]
            position_dbpm_weights = DBPM_N.loc[:, position]

            # Calculate the weighted sum for OBPM and DBPM
            obpm = np.dot(position_obpm_weights, row[o_play_attributes])  # Weighted sum for OBPM
            dbpm = np.dot(position_dbpm_weights, row[d_play_attributes])  # Weighted sum for DBPM

            # Accumulate the results for the player
            player_obpm += obpm
            player_dbpm += dbpm
        except KeyError:
            continue  # Skip the position if it's not found in OBPM/DBPM weights

    # Store the calculated values for the player
    obpm_values.append(player_obpm)
    dbpm_values.append(player_dbpm)

# Add the calculated OBPM and DBPM values to the dataframe
data_2022['OBPM'] = obpm_values
data_2022['DBPM'] = dbpm_values
```

Scatter Plots:

- Visualize how OBPM and DBPM correlate with player market value.
- Highlight players with balanced offensive and defensive contributions.

Interplay of Age, Value, OBPM, DBPM, and Potential

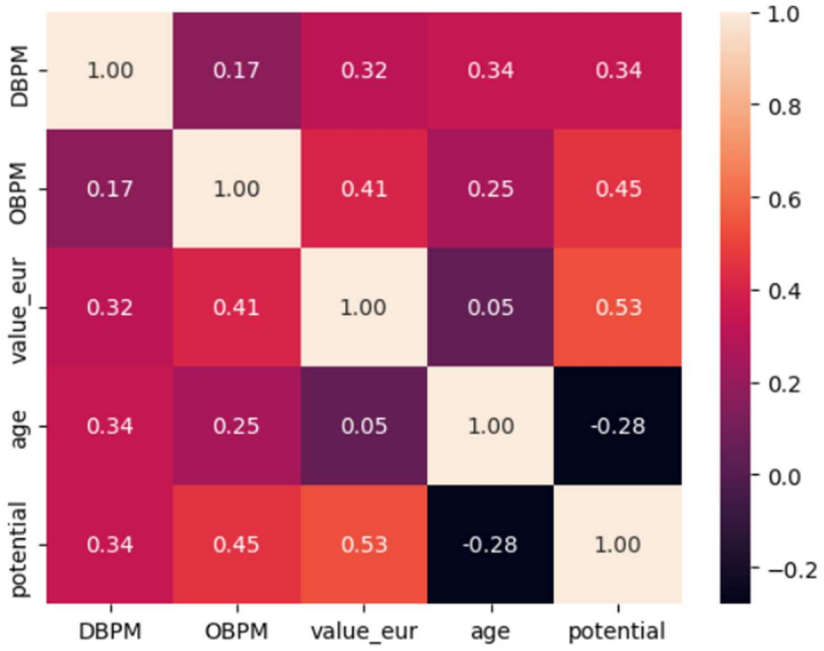


Figure 11: Target Variables Correlation Matrix

- **Age vs. Potential**: This shows a negative correlation, confirming that younger players tend to have higher potential, while older players exhibit declining prospects.
- **Age vs. Value (EUR)**: Younger players are often more highly valued, but this trend flattens or reverses older players with high current performance.
- **OBPM vs. DBPM**: A moderate positive correlation indicates that players with strong offensive contributions may also contribute defensively, though this varies by position.
- **Value (EUR) vs. OBPM/DBPM**: Offensive contributions (OBPM) have a stronger correlation with market value compared to defensive ones (DBPM), suggesting clubs prioritize attacking impact when valuing players.

	CAM	CB	CDM	CF	CM	LB	LM	LW	LWB	RB	RM	RW	RWB	ST
pace	0.10	0.16	0.09	0.13	0.09	0.17	0.13	0.14	0.16	0.16	0.12	0.13	0.16	0.12
shooting	0.27	0.22	0.26	0.28	0.26	0.23	0.25	0.28	0.22	0.22	0.26	0.28	0.20	0.32
dribbling	0.32	0.31	0.33	0.31	0.33	0.30	0.32	0.31	0.32	0.30	0.32	0.30	0.33	0.30
passing	0.30	0.31	0.33	0.28	0.33	0.31	0.29	0.28	0.30	0.31	0.30	0.28	0.31	0.27

	CAM	CB	CDM	CF	CM	LB	LM	LW	LWB	RB	RM	RW	RWB	ST
defending	0.31	0.53	0.53	0.35	0.46	0.47	0.21	0.22	0.44	0.49	0.25	0.27	0.43	0.32
physic	0.38	0.31	0.32	0.30	0.34	0.28	0.37	0.35	0.30	0.28	0.37	0.34	0.31	0.38
pace	0.30	0.17	0.15	0.35	0.20	0.24	0.42	0.43	0.26	0.23	0.38	0.39	0.26	0.30

Figure 12: OBPM vs. Market Value

Metrics for Valuation

1. **Offensive Box Plus-Minus (OBPM)**
    - Derived from: Pace, Shooting, Dribbling, Passing.
    - Reflects offensive contributions like creating scoring opportunities.
2. **Defensive Box Plus-Minus (DBPM)**
    - Derived from: Defending, Physic, Pace.
    - Measures defensive skills like tackling, intercepting, and physical duels.
3. **Age**
    - Important for determining potential and peak performance periods.

Model Workflow

1. **Data Split**:
    - **80% Training** and **20% Testing**.
2. **Features (X)**:
    - OBPM, DBPM, Age.
3. **Target (y)**:
    - Player Market Value.
4. **Position-Specific Weighting**:
    - **Forwards**: Higher weights for Shooting and Dribbling.
    - **Defenders**: Higher weights for Defending and Physic.

Justification for training on 2022 data -

Figure 13: Player Value across Clubs

We see player value varies, essentially increasing every year across clubs and nationalities. Also, updated data with most recent requirements of different positions in terms of our attributes will contribute differently to OBPM and DBPM calc compared to 2015 data.

In our project, we chose to train the model based on player positions because different positions demand unique skill sets that evolve with changing tactics in football. Over the years, the requirements for various positions have shifted significantly due to advancements in team strategies and playing styles. For example, in 2015, a Centre-Back (CB) primarily focused on defensive attributes like Defending and Physic. However, by 2022, modern tactical approaches often require Centre-Backs to have a higher Pace to counter fast-paced attacks and participate in high defensive lines.

These dynamic changes in positional roles make it crucial to base our model on the most recent data to accurately capture current trends. Therefore, we used 2022 data for training our model to predict 2023 player market values. This ensures our model reflects the most up-to-date positional demands, leading to more accurate and relevant player valuations. By doing so, we account for the nuances of evolving player roles, making our predictions more robust and aligned with contemporary football strategies.

Handling Multi-Position Players

In soccer, many players can effectively perform in multiple positions (e.g., **Lionel Messi** as a **Left Winger (LW)**, **Central Attacking Midfielder (CAM)**, or **Center Forward (CF)**). To account for this variability, we adopted the following strategy:

1. **Data Expansion**:
2. For each player, we created multiple entries corresponding to each position they can play.
3. For example, if a player is listed as **[LW, CAM, CF]**, we generated **three separate data points**:
    - **Player 1: [LW]**
    - **Player 2: [CAM]**
    - **Player 3: [CF]**
4. **Metric Calculation**:
    - We calculated **OBPM** and **DBPM** for each entry based on the position-specific weights.
5. **Selecting the Best Performance**:
    - After calculating OBPM and DBPM for each position, we selected the **best OBPM and DBPM values** across all the player's positions.
    - This approach helps us identify:
        - The **position where the player performed best** during the season.
        - The **highest possible base value** for the player, which can be valuable during contract negotiations or transfer evaluations.

Advantages of This Approach

1. **Performance Optimization**: By analyzing multiple positions, we ensure that players are evaluated based on their **optimal performance** rather than being constrained to a single role.
2. **Accurate Valuation**: This method provides a **more comprehensive valuation** by considering the versatility and adaptability of players, which are crucial traits in modern football.
3. **Strategic Insights**: Clubs and agents can use these insights to determine the **most valuable roles** for a player, aiding in transfer decisions and tactical planning.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer

# Prepare data
predictors = ['age', 'OBPM', 'DBPM']
target = 'value_eur'

X = data_short[predictors]
y = data_short[target]

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a regression model (example: Random Forest)
model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"R^2 Score: {r2*100:.2f}%")

```

```
R^2 Score: 85.22%
```

Figure 14: Model Accuracy

Model Performance Evaluation

Our Random Forest Regressor demonstrated strong performance in predicting player market values, achieving an accuracy of 85.22% on the test dataset. The model's reliability is evident from the close alignment between the original (actual) values and the predicted values for most players. This consistency highlights the effectiveness of using OBPM, DBPM, and age as predictive features for market valuation.

Results:

Model Performance

- **Accuracy**: The Random Forest model achieved **85.22% accuracy** on the test set.
- **Feature Importance**:
    - **OBPM**: The most influential feature in predicting market value.
    - **DBPM** and **Age** also contributed significantly.

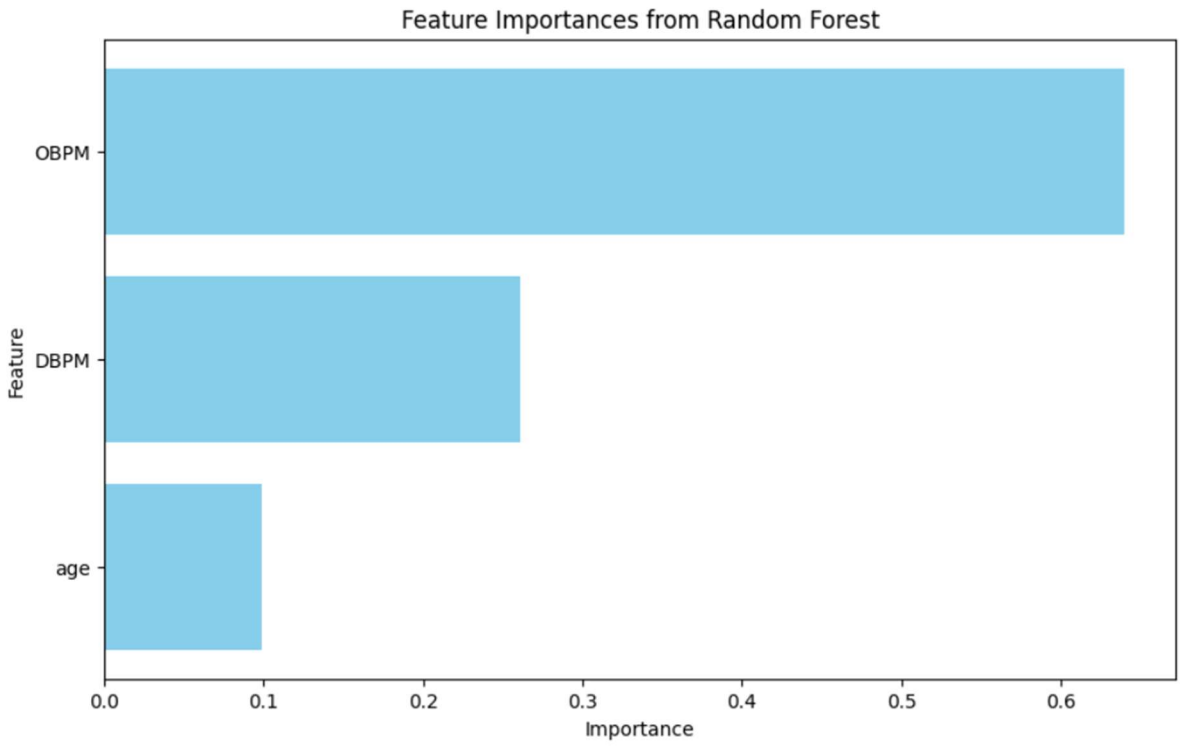


Figure 15: Feature Importance Chart

Visualization of Predictions

- **Dual Bar Plot**:
    - Compares real vs. predicted player values.
    - Highlights the model's accuracy and areas of deviation.

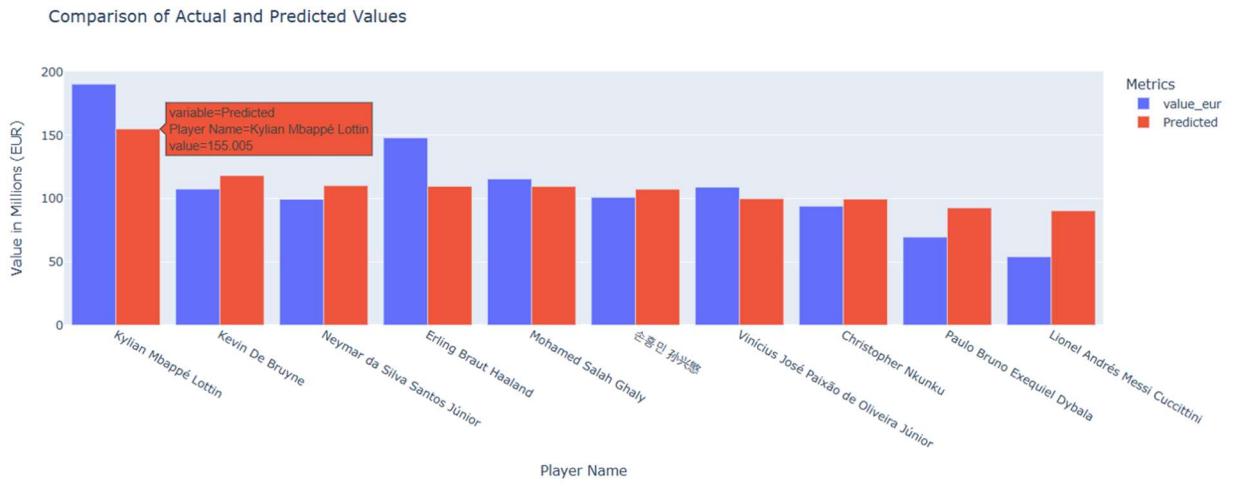


Figure 16: Real vs. Predicted Values

Case Studies: Notable Players

To further validate our model, we examined specific high-profile cases to understand where the model's predictions aligned or diverged from actual market valuations.

1. **Kylian Mbappé**
    - Observation: The actual market valuation for Mbappé was higher than our model's prediction.
    - Reason: His valuation was influenced by Real Madrid's willingness to pay a premium, reflecting market dynamics where elite clubs compete for star players, sometimes leading to overvaluation.
    - Insight: While our model provides an objective performance-based valuation, external factors like club demand and financial capacity can push valuations higher than predicted.
2. **Lionel Messi**
    - Observation: The actual market valuation for Messi was lower than our model's prediction.
    - Reason: His recent transfer to Inter Miami in the MLS saw his valuation dip due to the less competitive and less lucrative US market compared to European leagues.
    - Insight: This demonstrates that despite a player's high-performance metrics, market-specific factors (e.g., league strength, commercial potential) can lead to undervaluation.

Key Takeaways

- Model Strength: The model accurately captures the performance-driven market values for most players, validating its robustness.
- Contextual Factors: While our model provides a solid baseline, external factors such as club demand, league strength, and market dynamics can cause deviations.
- Strategic Use: Clubs and agents can use our model's valuations as a starting point for negotiations and to identify undervalued or overvalued players based on objective metrics.

This analysis demonstrates that our model is a reliable tool for player valuation while highlighting the importance of considering contextual market factors for comprehensive decision-making.

Conclusion

- **Accurate Player Valuation:** Our model predicts player value based primarily on their skills and age, ensuring a more objective and realistic valuation.
- **Mitigating Biases:** By excluding factors like leagues, nationality, and ethnicity, the model addresses common biases in current player valuations, such as overvaluation or undervaluation due to these external factors.
- **Objective Basis for Decision-Making:** This model provides a more reliable base valuation for players, allowing managers, club owners, and stakeholders to make more informed decisions on player acquisitions, purely based on performance attributes and age.
- **Recommendation for Use: We recommend using this system** as a core tool for evaluating player market value. While factors like player popularity can influence a club's decision for revenue purposes, our model ensures that the valuation process remains focused on skills and age, providing a fairer and unbiased starting point.
- **Practical Application:** Using the model's valuation, clubs can better determine how much to invest in players, aligning financial decisions with a player's inherent abilities and potential, rather than being swayed by external factors.