Data Analytic & Artificial Intelligence

Group 11

FIFA 22 Introduction

1. Game Overview:

- Published by EA sport.
- Release date: October 1, 2021.
- Available on PS4, PS5, Xbox, PC, and Nintendo Switch.

2. Game Features:

- Authentic football simulation experience.
- Includes both club and national team matches.

3. Database:

- The FIFA22 dataset provides in-depth information on thousands of players.
- Covers abilities, skills, and physical attributes.

FIFA 22 Players Dataset

Dataset Overview

1

- Number of players (rows): 19,239
- Different information (columns): 110

Type of information

2

- Player Information: Includes appearance, name, club information, salary, age, height and weight
- Player Ability Values: Includes ability values for each position and ability for each attribute.

Data Sources

3

Dedicated scouting teams and data analysts collect and analyze data on players' performance in the game.

Our Business Question

Our Business Question:
- How specific player attributes in this dataset impact their market value?

Our Aim:
- To predict the market value of players.
- To identify players who are undervalued or overvalued.

Why do we do?
- Football clubs face stiff competition and high expenses. Understanding and predicting player market values is essential for club leaders.
- We suggest creating a predictive model with FIFA 22 dataset and machine learning. This model will look at player attributes and performance, offering benchmarks for valuations to ease negotiations and spot players with mismatched market values.
- Help clubs gain a competitive advantage.

Methodology Steps

Step 1
Exploratory Data Analysis

Understand the distribution and relationships between player attributes

Step 2
Data Cleansing & Preprocessing

Prepare the data for modelling, ensuring it is clean and suitable for analysis

Step 3
Model Development & Evaluation

Development and evaluation of predictive accuracy models

Step 4
Model comparison & in-depth analysis

Compare the three models and identify the most appropriate model for in-depth study.

Step 5
Recommendation

Make business recommendations and policy recommendations based on your data analyses

Exploratory Data Analysis (EDA)

EDA aims to understand the dataset's structure, distributions of key variables, and initial insights into relationships between variables, mainly focusing on how age, overall rating and potential relate to market value.

Total Players

19,239 Players

Representing a wide range of skill levels, from emerging talents to world-class stars

Age

16-54 Years

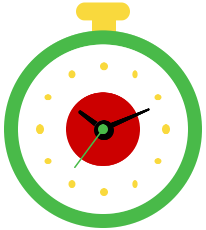
With median age of 25, indicating a mix of youth and experience

Overall Rating

47-93 Rating

With an average of 66, illustrating the broad spectrum of talent within the game

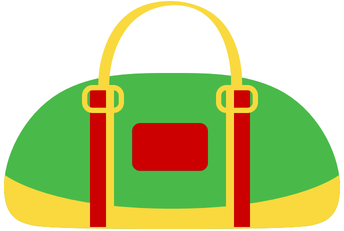
Potential



49-95 Rating

With an average of 71, show a promising future for many players

Market Value



€9,000- € 194,000,000

Highlight the economic disparities in player valuation

Distribution of Key Variables

Age

Distribution is skewed to the right, indicating a higher number of younger players

Overall Rating

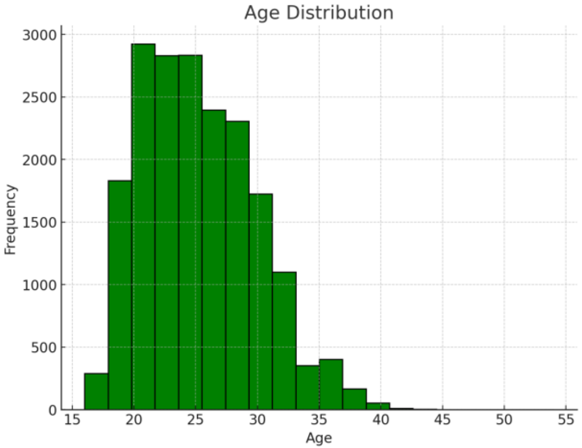
Distribution appears to be normally distributed with a peak around mid-60s to early 70s

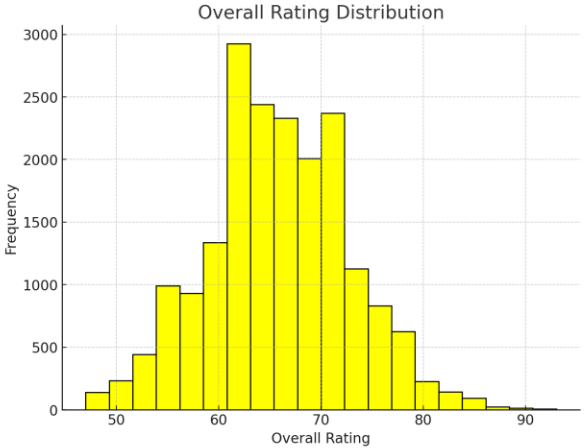
Potential

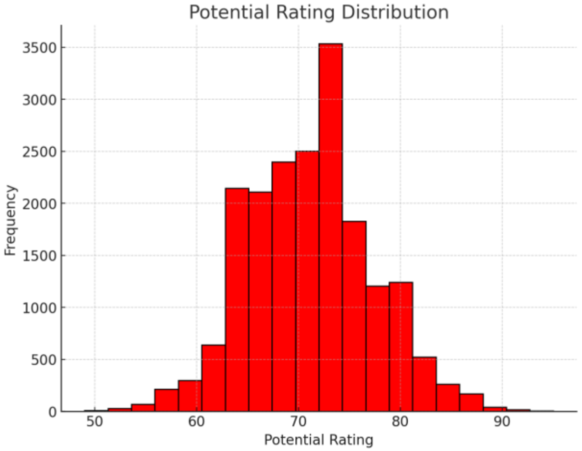
Distribution is slightly skewed to the left, indicating that more players have a higher potential rating

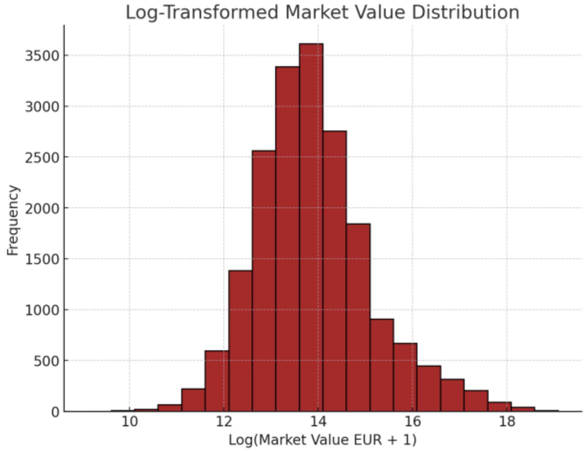
Market Value

The log-transformed distribution represents a large number of players valued at relatively low









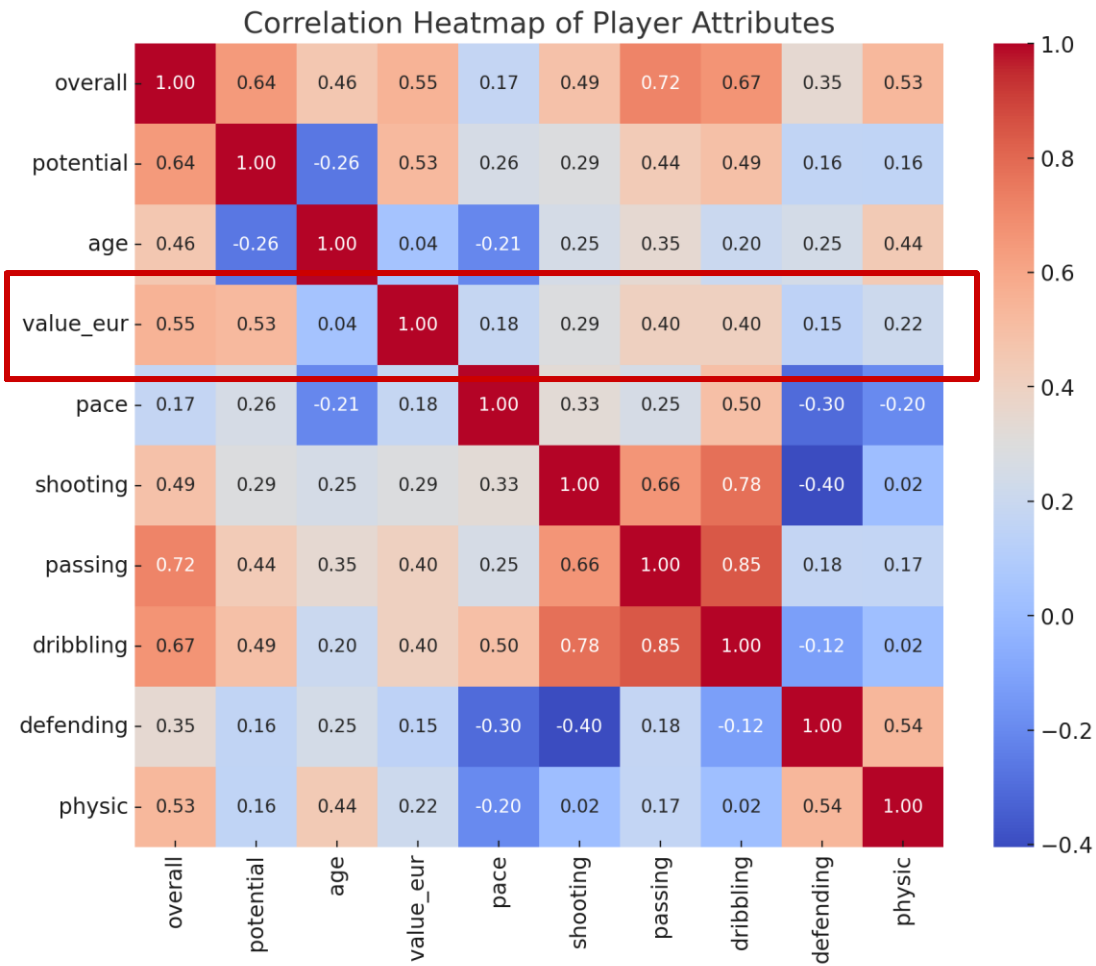
Relationship between Key Variables and Market Value

Age A player in their late 20s to early 30s might have higher market values, reflecting their peak playing abilities

Overall Rating A Clearer positive correlation between player's overall rating and their market value, indicating that performance and skill levels are significant factors in determining a player's market value

Potential Similar to the overall rating, there is a positive relationship between a player's potential and their market value, suggesting that clubs are willing to invest in future performance capabilities

Detailed Correlation Heatmap



Key Finding

The correlation suggests that market value is strongly influenced by a player's potential for future growth, current ability commanding higher market value.

Overall Rating and Market Value (0.55)

There's a strong positive correlation, indicating that players with high overall ratings tend to have higher market values.

Potential and Market Value (0.53)

The positive correlation between potential and market value indicates that players with higher potential ratings will likely have higher market values.

Data Cleansing & Preprocessing

Handle Missing Values

Remove missing values in the selected features

&

Split The Data

Splitting the data into training and testing sets

Select Relevant Feature

Including both player attributes and performance

Raw Data

Initial data	19239
Remove null/missing data	19167
Non-GK	17042
GK	2125

Handling missing values approach is to remove rows, as imputing them could introduce bias since our goal is to predict these values accurately.

- **Data Integrity:** imputing missing values may not accurately reflect the true attributes of players.
- **Impact on Model Performance:** imputation methods rely on assumptions that may not hold true across all variables.
- **Availability of Sufficient Data:** the dataset is large and comprehensive, removing a small percentage of rows with missing values will not significantly impact overall dataset size and ability to represent the population.
- **Simplicity:** Removing rows is a straightforward approach that makes the dataset-cleaning process simpler and more transparent.

GK

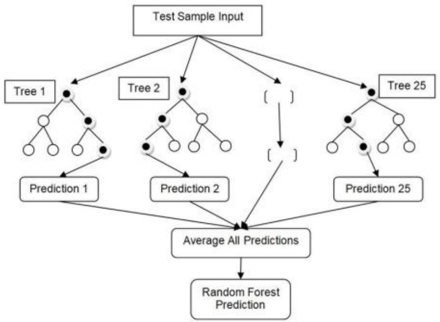
long_name
age
overall
potential
value_eur
wage_eur
player_positions
height_cm
goalkeeping_diving
goalkeeping_handling
goalkeeping_kicking
goalkeeping_positioning
goalkeeping_reflexes
goalkeeping_speed

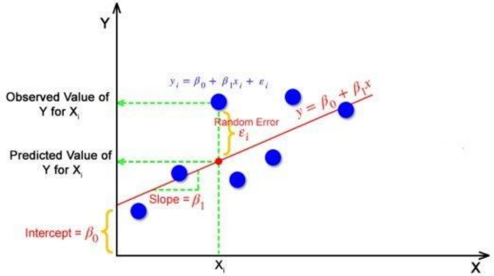
Non-GK

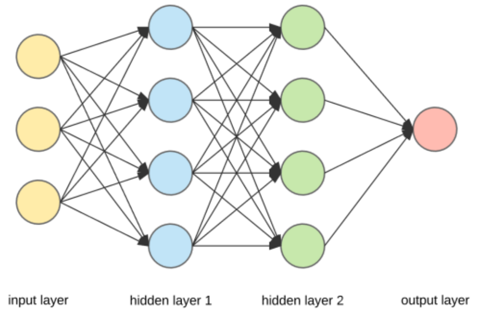
long_name
age
overall
potential
value_eur
wage_eur
player_positions
primary_position
position_category
height_cm
pace
shooting
passing
dribbling
defending
physic

Reason to Select Relevant Features:
- Avoid Multicollinearity
- Focusing on specific attributes
- Understanding underlying Value Drivers

Data Analytical Model and AI Methods

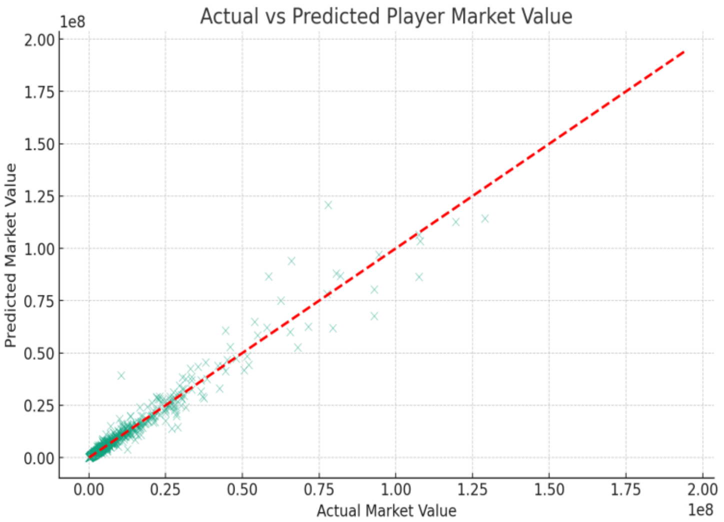


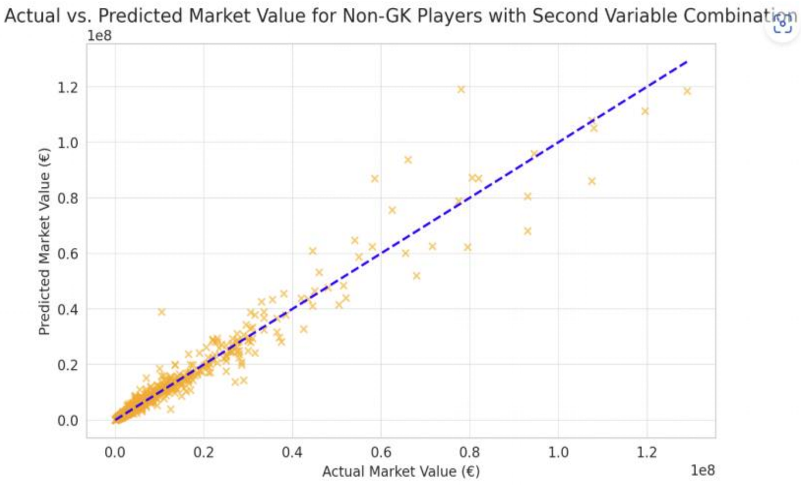




Random Forest	Linear Regression	Neural Network
Handling Non-linearity	Simplicity and Interpretability	Handling Complex Non-linear Relationship
Robustness to Overfitting	Baseline Model	Scalability and Performance

1. Random Forest Model





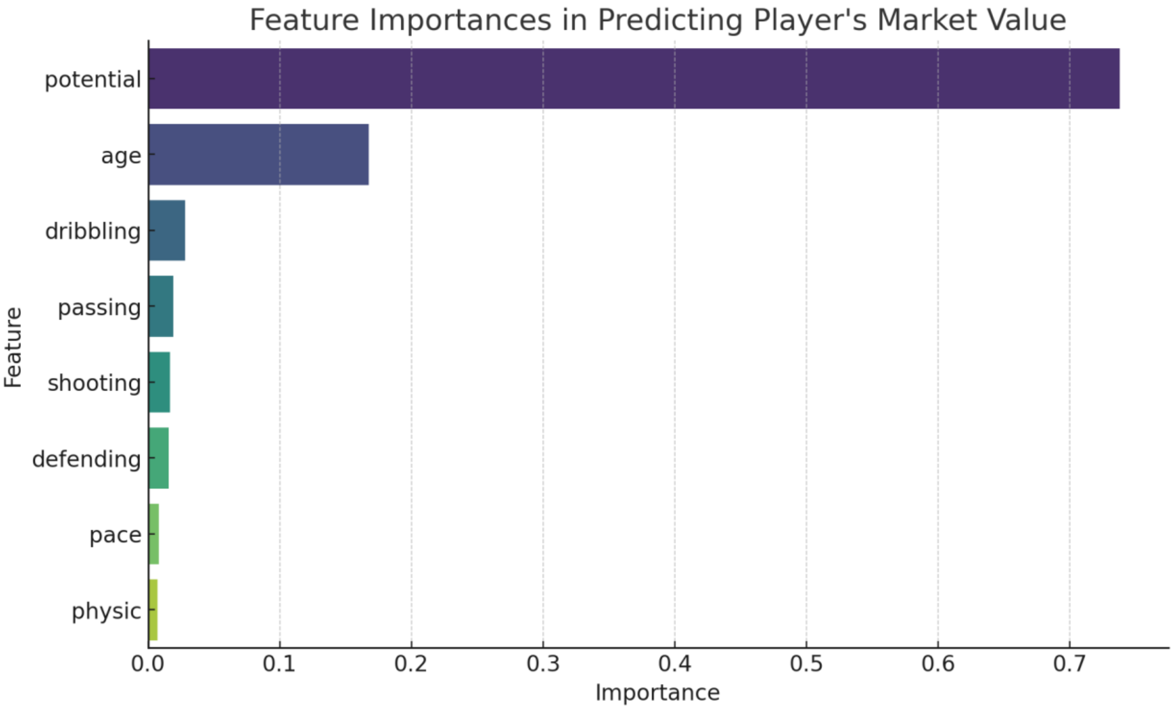
Summary

The result demonstrate that the Random Forest models are capable of predicting a player's market value with high degree of accuracy with the high score of R^2. This analysis underlines the importance of player attributes in determining their market values

Model Performance for Predicting Market Value

FIFA 22 Players (non-GK)	Mean Absolute Error (MAE)	407,196
	R-squared (R^2)	0.958
FIFA 22 Players (GK)	Mean Absolute Error (MAE)	193,949
	R-squared (R^2)	0.989

1. Random Forest Model

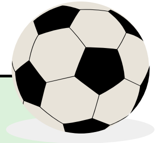


Feature Importance

The feature importance reveals how each attribute contributes to predicting a player's market value in professional football.

As shown, the 'Potential' and 'Age' are the most influential factors.

2. Linear Regression Model



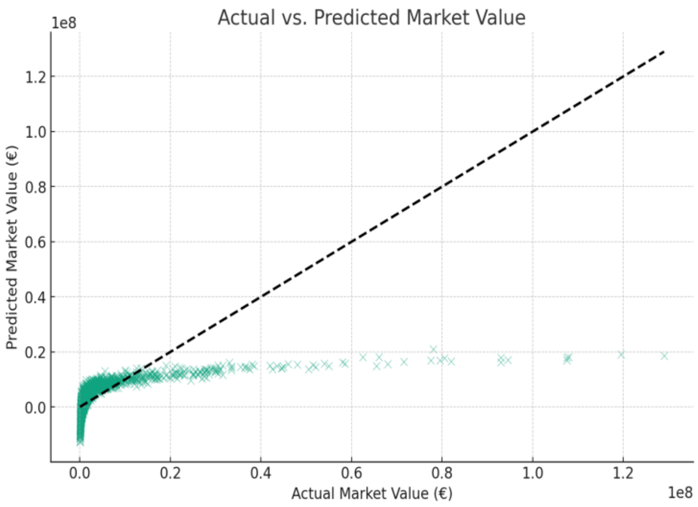
VIF Calculation	
Overall	10.15
Age	4.57
Potential	5.66
Pace	1.75
Shooting	4.93
Passing	6.20
Dribbling	8.08
Defending	3.75
Physical	2.40

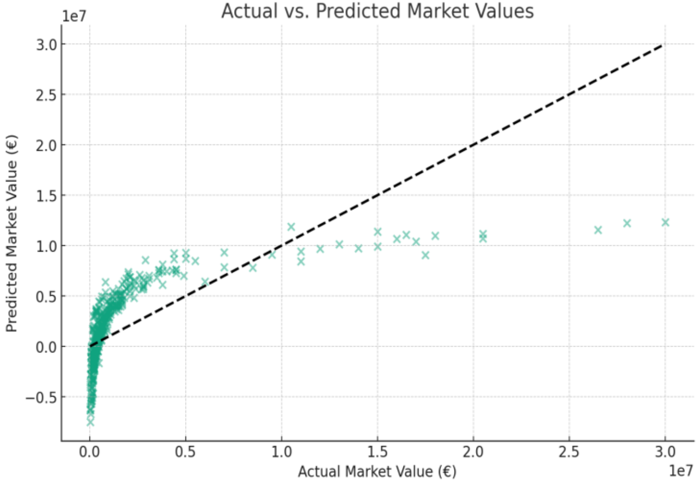
The **'overall'** rating, with a Variance Inflation Factor > 10 indicate notable multicollinearity. These results suggest that the overall rating is highly correlated with other player attributes.

In predictive modelling, high multicollinearity among independent variables can inflate the variance of regression coefficients, making the model less reliable.

Therefore, an **'overall'** rating is excluded as an approach to mitigate multicollinearity.

2. Linear Regression Model





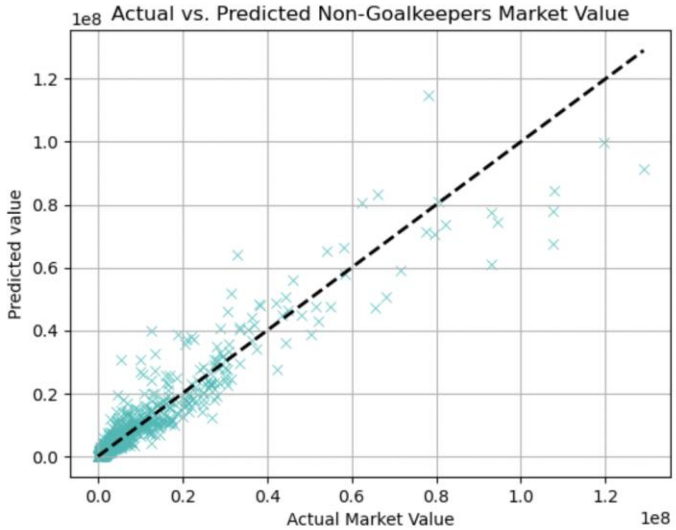
Summary

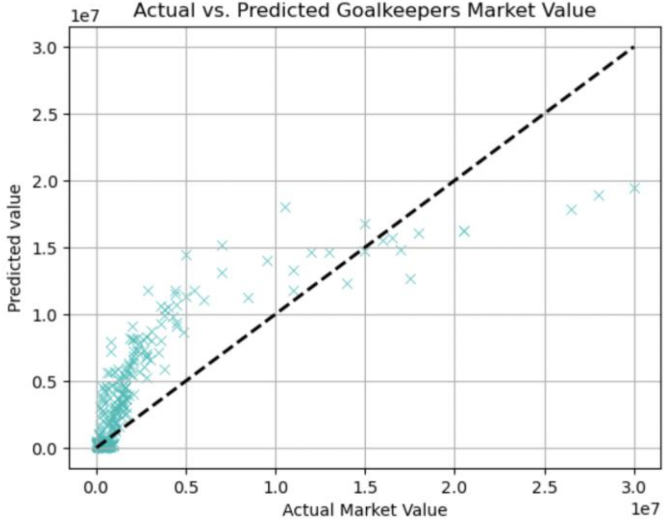
This model provides a foundational insight into how specified player attributes contribute to predicting a player's market value.

However, the R^2 values indicate that while the selected attributes have some predictive power, incorporate more complex modeling techniques can better capture the relationship between player attributes and their market value.

Model Performance for Predicting Market Value

FIFA 22 Players (non-GK)	Mean Absolute Error (MAE)	3,160,015
	R-squared (R^2)	0.334
FIFA 22 Players (GK)	Mean Absolute Error (MAE)	2,365,381
	R-squared (R^2)	0.3





3. Neural Network Model

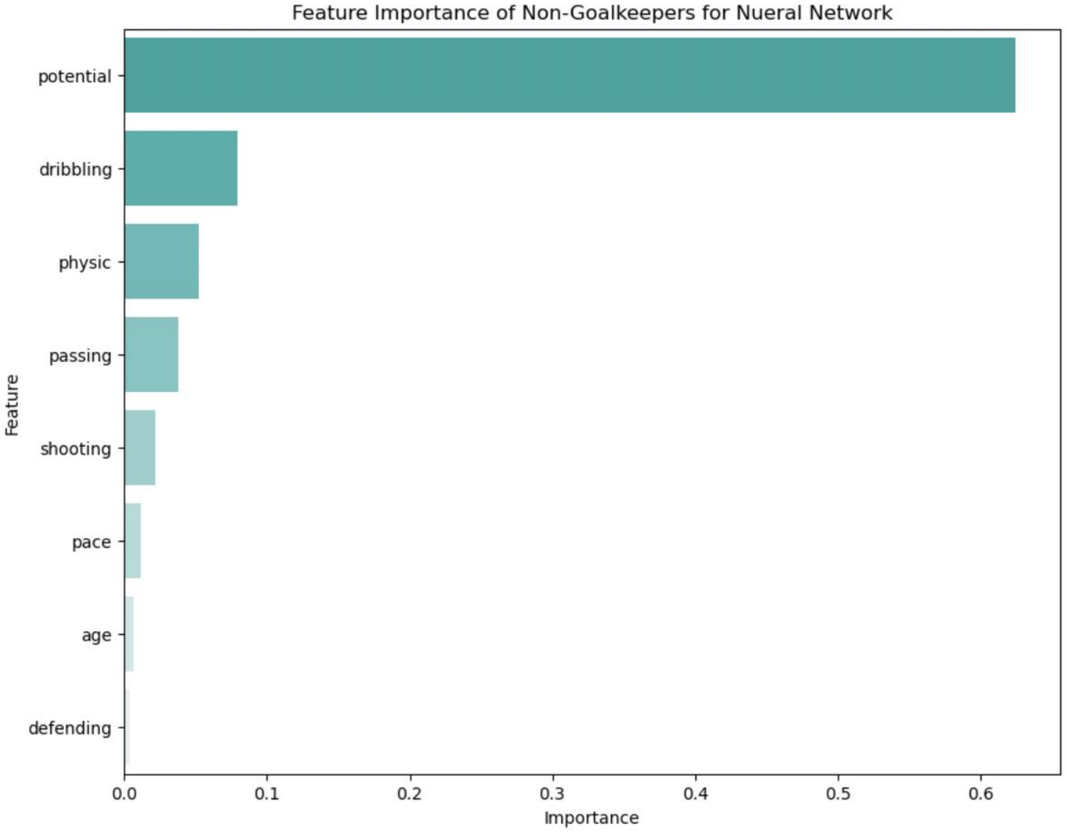
Summary

These provide a basic framework for a neural network regression model for predicting the 'value_eur' variable based on the selected features in the dataset.

Here, the R^2 is. 0.9 for non-goalkeepers, and 0.5385 for goalkeepers, implying that the model's predictive performance for goalkeeper market values is less accurate, potentially due to the unique attributes and valuation criteria associated with this position.

Model Performance for Predicting Market Value		
FIFA 22 Players (non-GK)	Mean Absolute Error (MAE)	903,053.5
	R-squared (R^2)	0.903
FIFA 22 Players (GK)	Mean Absolute Error (MAE)	1,451,827.6
	R-squared (R^2)	0.539

3. Neural Network Model

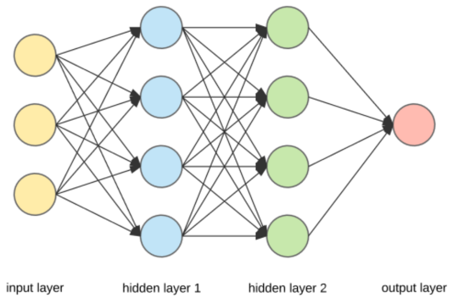


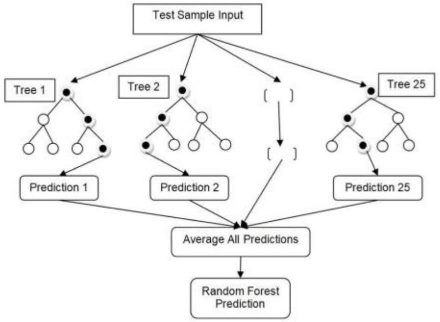
Feature Importance

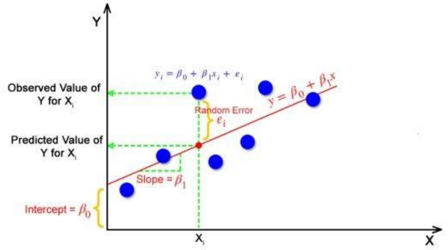
Here, from the importance of features in Neural Network, we can see while 'potential' is still top vital factor and 'dribbling' and 'physic' are more influent in this model, not 'age' as in Random Forest.

Understanding feature importance can help us to simply the model and optimize the prediction.

Data Analytical Model and AI Methods Result







2nd
Neural Network

1st
Random Forest

3rd
Linear Regression

	Neural Network	Random Forest	Linear Regression
MAE	903,054	**407,196**	3,160,015
R^2	0.903	**0.958**	0.334

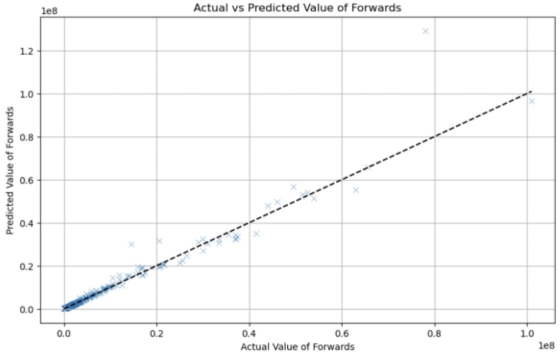
Random Forest In-Depth Analysis

For the outfield player, the data is divided into:	
Position	**Number of player**
Forward	3,666 players
Midfielder	7,009 players
Defender	6,366 players

Feature Importances for each position	
Position	**Top 5 out of 29 Features**
Forward	1. Movement Reaction 2. Ball Control 3. Dribbling 4. Short Passing 5. Finishing
Midfielder	1. Movement Reaction 2. Ball Control 3. Short Passing 4. Dribbling 5. Vision
Defender	1. Movement Reaction 2. Standing Tackle 3. Sliding Tackle 4. Sprint Speed 5. Stamina

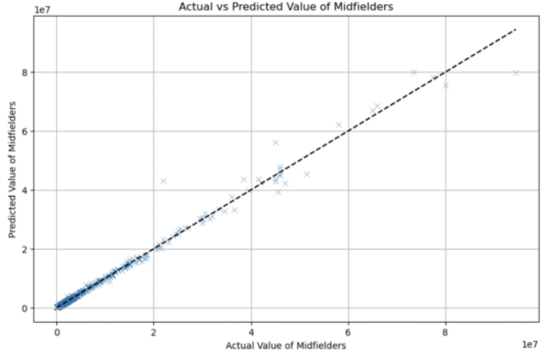
Forwards

R-squared value: 0.94
MAE: 323,399



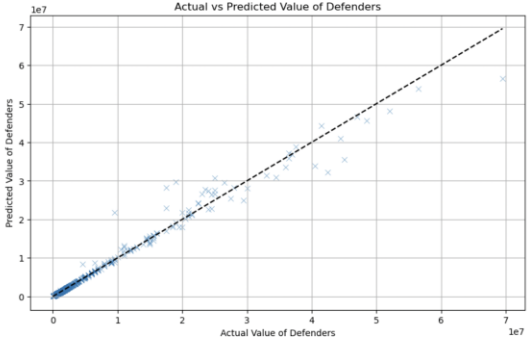
Midfielders

R-squared value: 0.99
MAE: 158,937



Defenders

R-squared value: 0.98
MAE: 169,558



Top 3 Undervalue Players

Forwards

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
Lionel Messi	Argentina	78,000,000	129,250,000	51,250,000
Zlatan Ibrahimović	Sweden	14,500,000	30,180,000	15,680,000
Dries Mertens	Belgium	20,500,000	31,735,000	11,235,000

Midfielders

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
David Silva	Spain	22,000,000	43,055,000	21,055,000
Sergio Busquets	Spain	45,000,000	56,195,000	11,195,000
Fábinho	Brazil	73,500,000	80,065,000	6,565,000

Defenders

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
Thiago Silva	Brazil	9,500,000	21,840,000	12,340,000
Juan Cuadrado	Colombia	19,000,000	29,740,000	10,740,000
Francesco Acerbi	Italy	17,500,000	28,195,000	10,695,000

Top 3 Overvalue Players

Forwards

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
Memphis Depay	Netherlands	63,000,000	55,390,000	7,610,000
Ben Yedder	France	41,500,000	35,035,000	6,465,000
Mauro Icardi	Argentina	37,000,000	32,290,000	4,710,000

Midfielders

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
Kai Havertz	Germany	94,500,000	79,875,000	14,625,000
Bukayo Saka	England	45,500,000	39,365,000	6,135,000
Carlos Soler	Spain	51,500,000	45,460,000	6,040,000

Defenders

Name	Country	Actual Value (Eur)	Predicted Value (Eur)	Value Difference
Achraf Hakimi	Morocco	69,500,000	56,565,000	12,935,000
Matthias Ginter	Germany	42,500,000	32,220,000	10,280,000
Stefan de Vrij	Netherlands	45,000,000	35,455,000	9,545,000

Business Problem Solution

Application of insights to solve the business problem

Our study using **random forest** model highlights that potential and age emerge as pivotal determinants shaping a player's market worth within the FIFA ecosystem.

Effect of Key Attributes	Accuracy on predicting the market value (Undervalued & Overvalued)	Predictive model as a base for Decision Making
Player's **potential** emerges as a significant predictor of market value. Players with **higher potential ratings** are likely to command **greater market value**, reflecting the **anticipation of future growth** and performance within the game. This underscores the importance of long-term strategic investments in promising young talents, as they hold the potential for substantial appreciation in market value over time. **Age** stands out as another crucial factor impacting player market value. Our analysis indicates that **younger players** generally fetch **higher market prices** compared to their older counterparts, **mirroring real-world trends** where youth is often associated with potential for development and longevity in performance. Consequently, clubs and players alike are inclined to prioritize younger talents in their transactions, seeking to maximize value and future returns on investments.	Identifying overvalued and undervalued players based on their attributes versus market values reveals **inefficiencies** in how the market values players. This discrepancy can arise from factors like **market hype, player nationality, club reputation**, or even **recent performances** in matches. From our analysis, we can help clubs and agents better understand where the market may be overvaluing or undervaluing certain types of players.	By integrating this predictive model (random forest) into decision-making frameworks, club can navigate **complexities** of the **transfer market** more effectively, making **strategic investments** in players that align with their **competitive and financial objectives** since this approach is a more rational and is an evidence-based pathway to building a successful and financially sustainable football ecosystem as **opposed** to relying solely on **market buzz and subjective opinions**.

Make business recommendations / policy recommendations based on data analyses

1

Data-Driven Player Valuation

Clubs and sports organizations should adopt similar AI and machine learning models with real life data for accurate player valuation. This approach can optimize transfer budget spending by identifying undervalued or overvalued players.

2

Educational Workshops

Conduct workshops for club executives and managers on the use of AI and data analytics in sports management, emphasizing the ethical use of data for player valuation.

3

Scouting Efficiency in videogame

Incorporate the random forest model to improve the scouting and career mode features, allowing players to uncover undervalued talent or negotiate more realistic transfers based on predicted market values.

Thank you