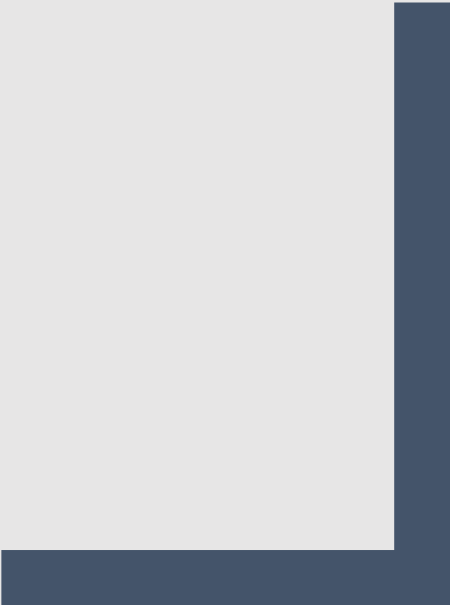


# **Evaluating the Market Value of FIFA players using Machine Learning**



## Description and Motivation of the Problem

In the world of football, a player's market valuation has gained substantial importance and interest after the 1995 Bosman ruling by the Court of Justice of the European Union, which fundamentally reshaped the dynamics of the transfer market (Europa.eu, 2024) by giving players' freedom of movement when their contract expires. In 2019, 7.35 billion dollars was paid by clubs to recruit players protected by employment contracts (MEN PROFESSIONAL FOOTBALL A REVIEW OF INTERNATIONAL FOOTBALL TRANSFERS WORLDWIDE, n.d.).

This shows that the valuation of a player has a significant impact in the world of football. From the club's perspective, this takes place for decision-making purposes. On the other hand, it is also helpful for the players' representatives to assess what value the club attaches to the athlete when negotiating his salary.

In Herm's et al. research (Herm, Callsen-Bracker and Kreis, 2014), they try to define what the market value in the professional football world is as "an estimate of the amount of money a club would be willing to pay to make [an] athlete sign a contract, independent of an actual transaction." "However, evaluating an individual's value within any kind of team – such as a soccer team – is a challenging task."

## Significance and Implications

In an era where player transfers, contract negotiations, and market valuations dominate headlines, the question of their valuation is of great interest for all parties (clubs, players and intermediaries) who need to assess a players worth in the transfer market. Understanding the intrinsic and extrinsic factors influencing a player's worth becomes paramount. Journalist Kaplan states that "the competition to provide soccer statistics reflects the level of interest in them." (Kaplan, 2010).

Traditionally, analysts gauge a player's market value using notational analysis, focusing on key performance indicators derived from statistical summaries of video footage and goal metrics. However, this approach is becoming outdated as machine learning evolves, offering more efficient ways to analyse intricate relationships.

Leveraging machine learning along with AI techniques to predict players' market value and determine the driving factors behind them not only aids clubs, agents, and stakeholders in crucial strategic decision-making but also provides a comprehensive perspective on talent valuation, market trends, and financial implications within the football ecosystem.

With all this in mind, there is a clear motivation to explore the relations between different features of football players and their market value - how these features affect their overall market value. Which are the most important features to affect the Market Value? And to use that knowledge to predict the expected market value of any given player.

## Dataset Overview

The analysis focuses on the FIFA 2024 dataset encompassing player attributes, performance metrics, and market values across diverse clubs and countries. It is a comprehensive record of players, capturing their skills, strengths and weaknesses within the global football market.

### Data Types and Relevance

- Discrete Variables: Player skill levels reflecting gameplay mechanics and roles.
- Continuous Variables: age, performance metrics (goals, assists, appearances), and other quantitative measures reflecting player valuation, form, and contributions.
- Categorical Variables: Countries and clubs

## Objective and Scope

The primary objective is to develop robust predictive models utilising artificial intelligence techniques to estimate players' current market values based on the dataset's features. By analysing, interpreting, and leveraging these variables, the analysis aims to identify

significant features that drive player valuations, allowing clubs and analysts to make informed, data-driven decisions within the dynamic of competitive football.

## Data Exploration and Pre-processing

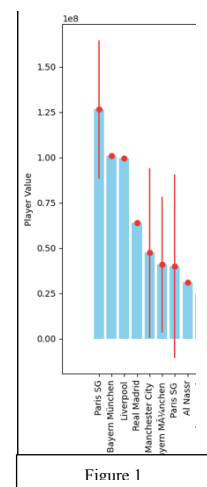
The main programming tools we will use for data exploration are Pandas, NumPy, matplotlib and seaborn for visualisation.

After loading our dataset in pandas, we would first like to identify the datatypes of the features we are working with and the amount of data in each feature. From this, we can see that we have a total of 40 columns. Of these columns, we have four columns with the datatype of 'object', among which is our target feature, 'value'.

The feature 'value' has the datatype of 'object' instead of 'int' due to its usage of a currency symbol, this will lead to complications when training the model, so it needs to be converted back to an integer type. This can be done using regex and casting

Other object type features, like 'country' and 'club', may need to be encoded prior to processing. From further investigation into these features (Figure 1), we see that there is a clear order to these values based on their mean value of players assigned to a given club or country. In light of this information, it would be best to encode them using label encoding.

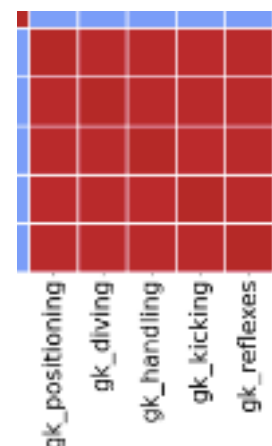
We can see that the marking feature has no value in it at all, so this feature can be dropped as it has no effect on the resulting market value of players. Along with that, we are also able to drop the feature 'player', as it is just the names of the players and does not have any relation to the target value. In addition to this, we can identify three counts of duplicate records, so we will remove them to avoid issues arising when building the models.



## Multicollinearity

We can further investigate this dataset to identify correlations between the target and input features as well as across input the features themselves. The former gives a general understanding as to which features are most significant in determining the target; this can prove to be a useful guide when building or improving our model, as it gives a basis for some feature engineering. The latter, on the other hand, gives an important insight into multicollinearity.

Multicollinearity refers to the situation where two or more features in a regression or predictive modelling context are highly correlated. You can see a subsection of a generated heat map that shows correlations between every feature. (The full heat map is viewable at the github link.) In this subsection, we are looking specifically at all goalkeeper-related features and as evident by the colour of the cells, these features are extremely high in correlation to each other. When multiple features have high correlation between each other, it is difficult to identify the individual effects that each feature will have on the target. To resolve this, we can make use of a technique called Principle Component Analysis to mitigate multicollinearity by reducing the number of features while retaining most of the original data's variance. This is achieved by generating orthogonal principal components that capture the maximum variance in the data.



## Summary of Methodologies Used

### Support Vector Regression (SVR)

SVR is a machine learning technique used for regression tasks that focuses on fitting a hyperplane that minimizes the error between predicted and actual values in a high-dimensional space, leveraging kernel functions for non-linear transformations. It is effective in capturing non-linear relationships whilst also robust to outliers. With proper parameter tuning, this could end up having the best performance. However, it will lack interpretability and may end up being quite computationally intensive.

### Decision Trees

Decision trees partition the feature space into regions based on a series of hierarchical rules, allowing for classification or regression based on attribute values. We will only be using regression trees. Decision trees are a popular pick for simplicity as it offers easy interpretability and insight into feature importance. But it does tend to be prone to overfitting and may struggle with capturing complex relationships in the FIFA datasets.

### Random Forest

Random Forest is an ensemble learning method that aggregates multiple decision trees to enhance predictive accuracy and mitigate overfitting and so, it is quite capable of providing reliable data. However, this does mean that the bagging technique it utilises can introduce complexity, which may reduce interpretability. This method is almost in the middle ground between the two previously mentioned methods, which implies it will be the best method for our specific problem.

### Metrics Used

The primary evaluation metrics to assess and compare the performance of all 3 models and their iterative versions were Mean Squared Error (MSE) and R<sup>2</sup> (R-squared). The choice of MSE provides insight into the average squared differences between the predicted values and the actual values, emphasising the importance of predictive accuracy. Concurrently, R<sup>2</sup> offers a measure of the proportion of variance explained by the models, aiding in understanding their efficacy in capturing underlying relationships. The range of R-square is normally 0 to 1, and the closer the value of R-square is to 1, the better the model fits.

### Data Splitting

The dataset underwent a division into features (X) and the target variable (y), with 'value' denoting the market value of players. To facilitate robust model training and assessment, a train-test split was executed with a ratio of 70% for training and 30% for testing. This division ensures that the model is exposed to a substantial amount of data during training, enabling it to learn patterns and relationships effectively. The training set is further split again, with the same ratio, but this time 30% is for validation. This nested split further allows for the optimisation of model hyperparameters on the training-validation subset while maintaining a separate dataset for final unbiased evaluation.

### Learning Curve Analysis

A learning curve was constructed to provide a visual representation of the model's behaviour across varying training set sizes. The mean and standard deviation of the negative Mean Squared Error (MSE) for both the training and cross-validation sets were computed and graphically depicted against the changing sizes of the training set. This analysis allowed for a comprehensive understanding of how the model's predictive performance evolves with different amounts of training data, providing insights into potential overfitting or underfitting.

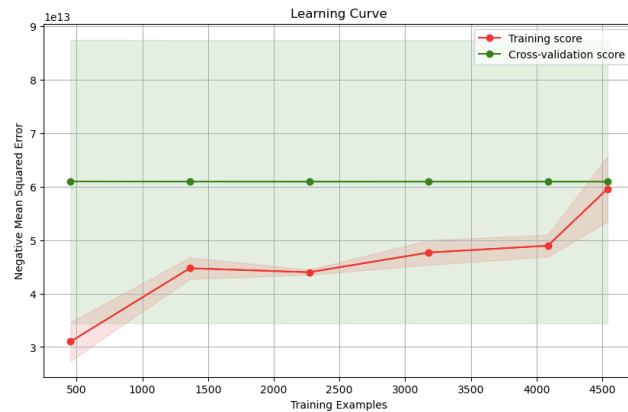
## Results and Evaluation

### SVR

Originally, for this model, like for all others, we label encoded the categorical features in the dataset (country and club). However, whilst fine-tuning this model, I found that it actually performs much better if these features were completely omitted.

### Learning Curve Analysis

It is worth noting that this model had a rather interesting learning curve. The stable cross-validation score indicates that the model's ability to generalise to unseen data is not improving despite increasing training performance. This could be caused by the SVR model becoming too complex where it captures both the underlying patterns in the data and the noise in the training set. To resolve this, we can make use of hyperparameter-tuning.



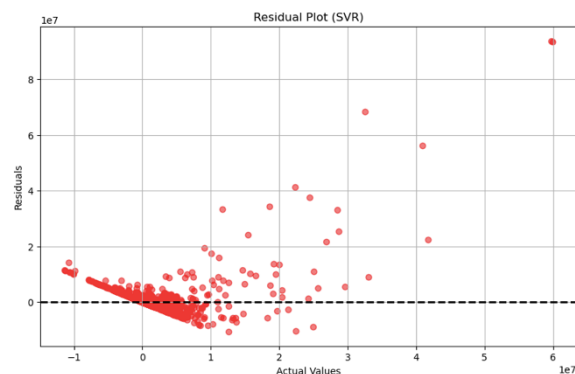
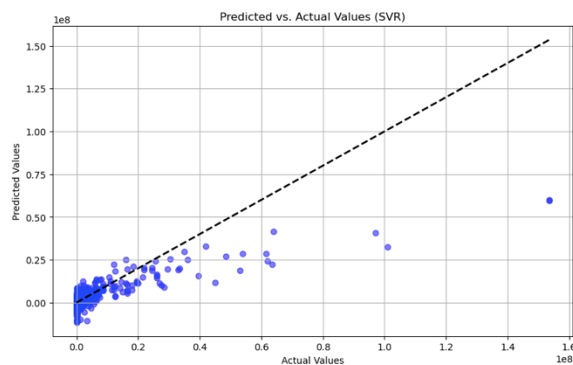
### Outcome

After hyperparameter-tuning, the best Mean Squared Error (MSE) and R-squared, have significantly improved:

Mean Squared Error (MSE): 54002818335301.51  
R-squared Score: -0.09552641887621993

R<sup>2</sup> Score: 0.6122206909893722  
Mean Squared Error: 36571888521032.23

Using the generated predictions, we can compare the predicted market value and true market value of all players to plot the graphs of true market values against predicted market values and a residual plot to better visualise the accuracy of this model:

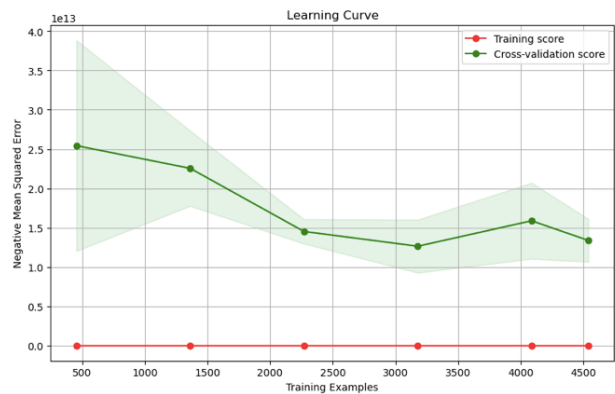


The closer the blue points are to the dashed line, the closer the predicted market values are to the true market value, meaning that if the blue points are on the red line, the predicted values are equal to the true values. Unfortunately, even with the improvements, it is evident that many predicted results are indeed not accurate; in fact, the points are further from the dashed line when the expected market value is larger, which portrays this model has a larger variance when trying to estimate large market values. Furthermore, due to the fit of the SVR model, some players' predicted market values are even negative, so with this in mind and the difficulty in interpreting this model, we concluded that this model is not significantly ideal for predicting and identifying the driving factors behind a player's market value.

## Decision Tree

### Learning Curve Analysis.

The plot shown demonstrates one of the many risks discussed beforehand of utilising decision trees, which is their proneness to overfitting to the training set. Although the cross-validation score initially approaches 0, which signifies that the model is improving, you can see that it consistently performs perfectly in predicting values against the training set. This implies that the model is not generalising the data well as the training set increases. This could be the reason why the cross-validation score fluctuates.

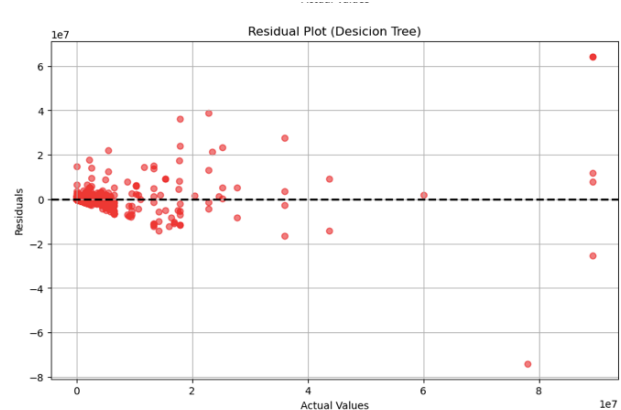
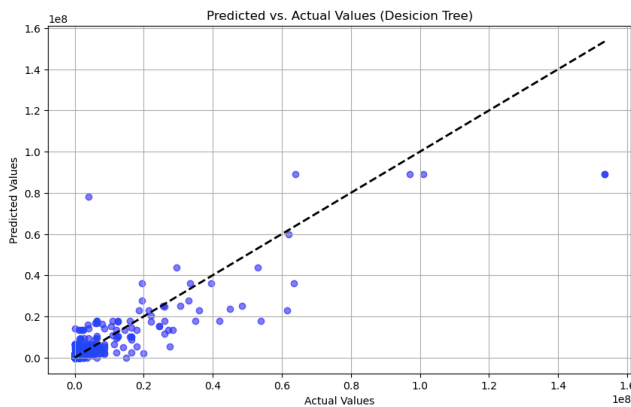


### Outcome

After rigorous tuning of the hyperparameters, we only managed to improve the model's R-Squared score but also increased the Mean Squared Error:

Mean Squared Error (MSE): 15005834109694.66  
R-squared Score: 0.6955846340764387

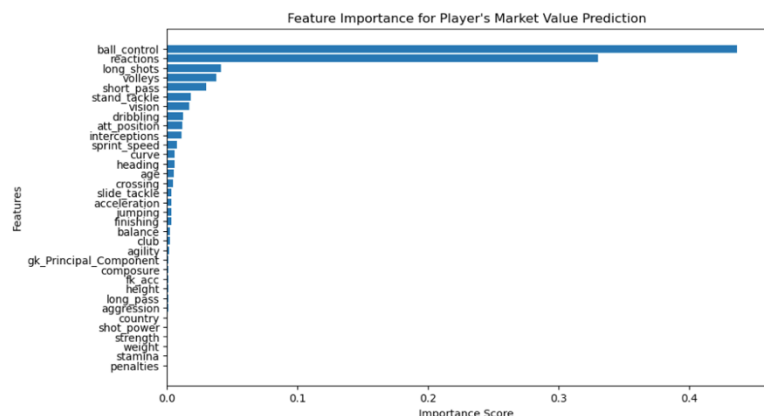
R<sup>2</sup> Score: 0.7602079719445927  
Mean Squared Error: 22615047049955.51



Regardless of the increase in the Mean Squared Error, this model still outperforms that of the SVR model. This model greatly benefited from the usage of Principle Component Analysis (PCA) on the goalkeeper fields, making its Mean Squared Error score than that of the SVR model and its R-squared score a large improvement from the previous. These improvements from the previous model are reflected in the plots, where you can see that the blue points, in general, align themselves closer to the dashed line. There are also no negative values, making this overall a much better model to use than our first one.

### Feature Importance

With this model generated and working to an acceptable standard, we can now grab the feature importance to identify the driving factors behind a player's market value. We can see that the most significant features are a player's ball control and their reactions, followed distantly behind by shots, volleys and short passes.

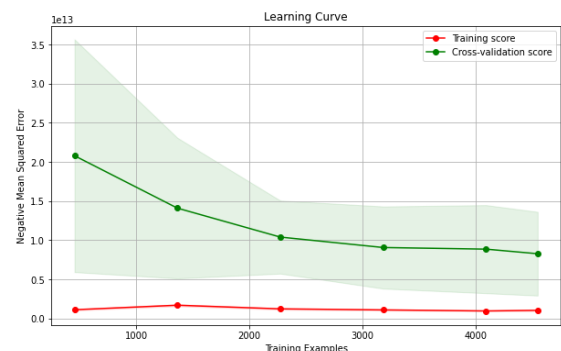


## Random Forest Regression

The Random Forest Regressor was selected as the model of choice due to its inherent capacity to effectively handle non-linear relationships within the dataset and its ability to discern feature importance, making it well-suited for the task of predicting player market values. The model was trained using the designated training set and subsequently tested using the dedicated test set. Following this, predictions were generated on the test set, and key performance metrics, Mean Squared Error (MSE) and R-squared, were computed.

### Learning Curve Analysis.

Analysing this graph, it can be observed that the cross-validation scores approach 0 as the training size increases, which is expected as the model benefits from more data. It indicates improved model accuracy and predictive power. However, the training scores show a more complex pattern. Initially, the indicating scores decrease, indicating improved generalisation, but at larger training sizes, they start to plateau or even increase. This suggests that the model may be overfitting or encountering difficulties in generalising to new data when the training set becomes too large.



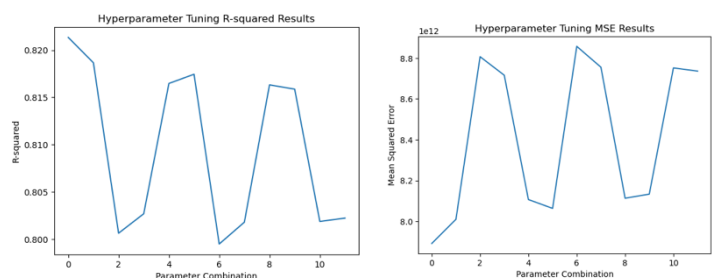
### Hyperparameter Tuning

#### Overall Trends:

Generally, as the max\_depth increases, the model tends to perform better, but there is a risk of overfitting. Lower values of min\_samples\_split and min\_samples\_leaf seem to contribute to better performance, but setting them too low might lead to overfitting.

#### Best Performances:

The combination with max\_depth = 20, min\_samples\_split = 2, and min\_samples\_leaf = 1 seems to have the lowest MSE and a high R2, indicating good predictive performance.



### Outcome

The model exhibited a lower Mean Squared Error than previous models and a higher R-squared score of 0.8996, indicative of its capability to account for a substantial proportion of the variance in the market values of football players.

Mean Squared Error: 7680617171183.443  
R-squared Score: 0.826115161310985

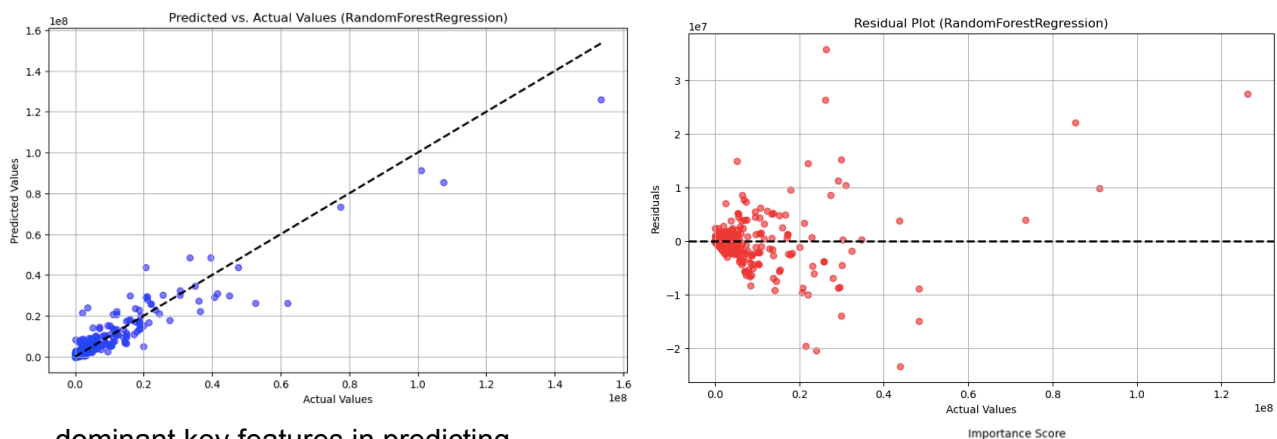
R<sup>2</sup> Score: 0.8996248071794241  
Mean Squared Error: 7214695831694.318

Visualisations, including a scatter plot of actual vs. predicted values and a residual plot, were created to assess the model's predictive capabilities. The scores achieved by this model are the most successful of all models. This is also conveyed in the plots by the closeness of the blue points to the red line, suggesting it has the least variance in its predictions compared to all other models. This is expected as this is an ensemble learning method which mitigates overfitting, allowing this model to achieve better results than a simpler Decision Tree method. This model is, therefore, the most ideal to use for our problem.

### Feature Importance

By grabbing the feature importance from the Random Forest model, we see a similarity between this model and the decision tree model, where reactions and ball control are the





dominant key features in predicting a player's market value. However, this model then considers composure and dribbling to be the next most significant instead of volleys and long shots like the decision tree model does.

It is important to remember that during the exploratory data analysis stage, we created a correlation heatmap between the input features and the target (a subsection of this is viewable on the right). The heatmap itself aligns closer to the feature importance ranking established by the Random Forest regression model.

## Evaluation Summary

The evaluation of machine learning models for predicting FIFA player market values provided insightful perspectives on their efficacy, limitations, and applicability within the football ecosystem.

ball_control	- 0.28
dribbling	- 0.25
slide_tackle	- 0.075
stand_tackle	- 0.094
aggression	- 0.18
reactions	- 0.5
att_position	- 0.25
interceptions	- 0.11
vision	- 0.34
composure	- 0.39

Starting with Support Vector Regression (SVR), the model initially demonstrated promising performance metrics, however, it encountered challenges in handling complexity and interpretability. The learning curve analysis and hyperparameter tuning revealed the SVR's constrained generalisation capabilities; notably, the model's computational intensity posed significant challenges during tuning, leading to inefficiencies which necessitate future improvement. Despite these limitations, insights into the SVR's intricacies allow us to learn the need for rigorous optimisation, feature engineering, and computational resources to enhance predictive robustness and mitigate complexities.

In contrast, Decision Trees offered a simple approach, providing interpretability and insights into feature importance. The model's susceptibility to overfitting, as evidenced by fluctuating cross-validation scores, highlighted the importance of rigorous hyperparameter tuning and feature engineering to balance complexity and performance effectively. Moreover, the Decision Tree's reliance on specific attributes highlighted the significance of comprehensive data pre-processing and domain expertise to discern meaningful patterns and enhance model reliability.

Finally, as expected, the Random Forest Regression model emerged as the best model, leveraging ensemble learning to mitigate overfitting and enhance predictive accuracy. With a far higher R-squared score of 0.8996, the model effectively captured non-linear relationships and discerned key features influencing player valuations. Despite its computational intensity and reduced interpretability compared to individual Decision Trees, the Random Forest's superior performance and minimal variance in predictions underscored its potential as a strategic tool for stakeholders in navigating player transfers, contract negotiations, and market dynamics.

In summary, the Random Forest Regression model demonstrated better predictive capabilities out of the explored models. However, it's crucial to acknowledge that each model has its strengths and weaknesses. The difficulty of achieving perfection emphasises the



ongoing need for refinement, validation, and collaboration with domain experts to navigate the evolving complexities and biases in the football world.

Moreover, there are many limitations to consider. For instance, the dataset used does not account for a player's popularity, which could influence their market value. Exploring such factors alongside domain experts is essential for a more comprehensive understanding of player valuation in the football industry. Future research could delve into hybrid approaches, leveraging multiple datasets to capture a more holistic understanding of player valuation, including factors like popularity, which may significantly influence market value but remain unexplored in the current dataset.

## References

Europa.eu. (2024). *EUR-Lex - 61993CJ0415 - EN - EUR-Lex*. [online] Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:61993CJ0415> [Accessed 1 Jan. 2024].

MEN PROFESSIONAL FOOTBALL A REVIEW OF INTERNATIONAL FOOTBALL TRANSFERS WORLDWIDE. (n.d.). Available at: <https://digitalhub.fifa.com/m/248987d86f2b9955/original/x2wrqjstwjoiainncnod-pdf.pdf>.

Herm, S., Callsen-Bracker, H.-M. and Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review*, 17(4), pp.484–492. doi:<https://doi.org/10.1016/j.smr.2013.12.006>.

Kaplan, T. (2010). When It Comes to Stats, Soccer Seldom Counts. *The New York Times*. [online] 8 Jul. Available at: <https://www.nytimes.com/2010/07/09/sports/soccer/09soccerstats.html>.