# Striker Market Valuation in Football : A Regression Analysis using FIFA 22 Data

F216622

Institute of Sport Business, Loughborough University

LLP315 Sport Business Statistics and Analytics

Dr. Daniel Read

## Executive Summary

Accurately estimating a football player's market value is an essential task in the process of player transfers that is traditionally performed internally at clubs based on domain knowledge. The use of data-driven analytical processes for this estimation has been scarcely used until recently and is limited to big clubs with strong finances and technological capabilities. This report aims to objectively define a method for predicting the market value of football strikers using data collected from the football simulator game, FIFA 22. Relevant independent variables were selected based on a literature review and used in a multiple linear regression to identify a significant relationship with the market value of a striker. The results of the study illustrate the viability of using game data for cost effective market value estimations that can be used as a baseline in real-life transfer negotiations.

## Introduction

A football player's market value is described as the most likely estimate of the player's transfer fees that an interested club is willing to pay (Muller et al., 2017). Traditionally, the market value of a professional football player has either been estimated internally by football clubs or by sports journalists. However, these valuations suffer from internal biases, inaccuracies, and a lack of transparency in their calculations (Al-Asadi & Tasdemir, 2022).

Muller et al. (2017) highlighted that the use of data-driven methodologies for player valuation has been scarcely used in the sport of football until recently. Crowdsourcing using publicly available data has led to the emergence of reliable estimations of a player's value. Al-Asadi and Tasdemir (2022) describe how expert-based estimations can be complemented with the use of data analytics.

Rich data that accurately describes the characteristics of a player in detail is not easy to collect and is usually limited to teams with strong finances and unrestricted technological access (Cotta, 2016). The absence of a cheap, accessible, widely available large-scale dataset of player statistics makes it difficult for smaller clubs and interested individuals with limited resources, to accurately derive a player's value on their own.

The objective of this report is to identify a data-driven process that can effectively quantify a football player's market value using publicly available data.

A review of relevant literature has shown that since 2014, researchers and football clubs have explored the use of football simulator video games as an alternate source of player data. Significantly accurate results were obtained by using game data from EA Sports FIFA for various machine learning projects (Al-Asadi & Tasdemir, 2022). The use of FIFA game data to identify the influence of individual player attributes on their market value has also been studied by Singh and Lamba (2019), and Behravan and Razavi (2021).

The aim of this report is to identify and establish a relationship between the market value of a football striker and their individual attributes that are obtained from the FIFA game data. A study of relevant academic theory has shown that this can be achieved with hypothesis testing of the significance of parametric models. Skinner et al. (2015) describes a step-by-step process to establish the null and alternate hypothesis, interpret the correlation coefficients of related variables, and accurately establish the significance of the chosen model using the p-values.

Miller (2016) highlights the importance of choosing the right set of variables when defining the parametric model, by explaining the effects of having too many, or too few variables. The accuracy and reliability of a regression model is subject to satisfying certain base assumptions of the chosen dataset, which can be established by conducting various tests using the SPSS software while creating the model. Thrane (2019) explains how to interpret the effects of the R-squared value, multicollinearity, homoscedasticity, normality, etc. to validate the performance and reliability of the regression model.

## Method

The data that shall be used to predict the market value of a football player is collected from the EA Sports FIFA 22 video game. This data set features the attributes and performance ratings of over 19,000 real life professional football players from across the planet. Each player is assigned a unique set of over 100 attributes such as passing score, shooting score, dribbling score etc. that describes their performance. The developer of the game, EA Sports estimates the attribute scores for each player by employing a wide network of scouts and analysts who watch real-life video footage and attend live games.

The attributes of a player are subject to change between each annual installment of the EA Sports FIFA game, based on how they perform in real life. This dataset was uploaded to Kaggle by Stefano Leone, who scraped the information from the website sofifa.com and saved it in the .CSV format.

The dataset initially consisted of 19,239 individual players, but as the scope of this analysis is limited to only players who operate as strikers, the dataset was filtered down to 3,398 players. The dependent variable chosen for this report is the market value of the player in Euros which is denoted by the variable *value_eur* in the FIFA 22 dataset. Based on existing literature on the available domain knowledge, only a few select independent variables were identified for the analysis due to their frequent appearance across the literature (Al-Asadi & Tasdemir, 2022).

Carmichael and Thomas (1993) described the relationship between a player's market value and their *age*, based on their potential for growth and experience. They noted that after a certain age, the player's ability begins to decline, which affects their market value.

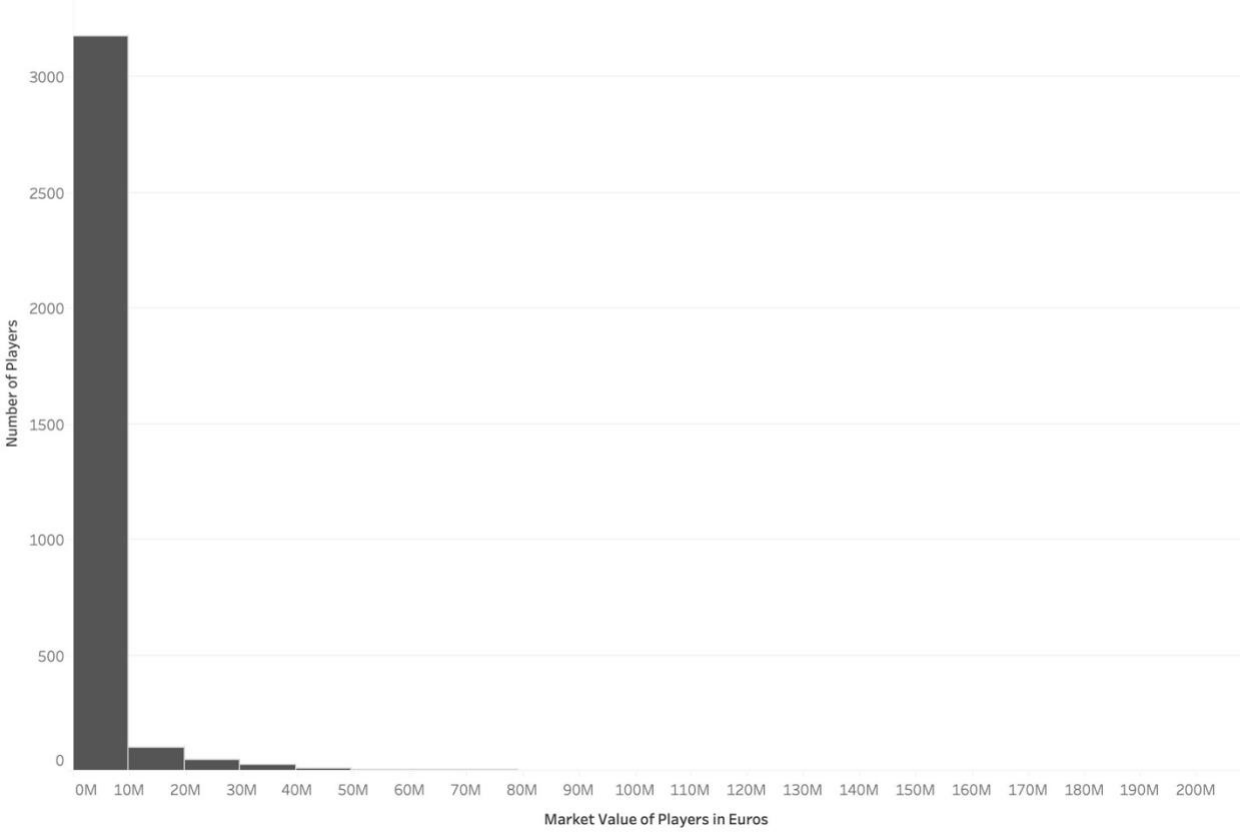
The *international reputation* of a player was found to be statistically significant for determining their market value in a study conducted by Al-Asadi and Tasdemir (2022). The quality of a player is noted to be supported by their international recognition at both the youth and adult levels by Gerrard (2001).

Using clustering methods to analyze the market values of players across all positions, Behravan and Razavi (2021) found the attributes of a striker's *shooting*, *dribbling*, and *physicality* skills to be important for determining their market value. Studies conducted by Muller et al. (2017), and Tosato and Wu (2018) supported the role of a striker's physicality in influencing their probability of scoring goals.

To establish a relationship between the dependent variable (*value_eur*) and the independent variables (*age, international_reputation, shooting, dribbling, physic*), the analysis chosen is a multiple linear regression, where the influence of all the independent variables is accounted for when predicting the dependent variable (Skinner et al., 2015). This analysis can also explain the variance caused in the dependent variable by each individual dependent variable, along with their significance of predicting power.

**Figure 1**

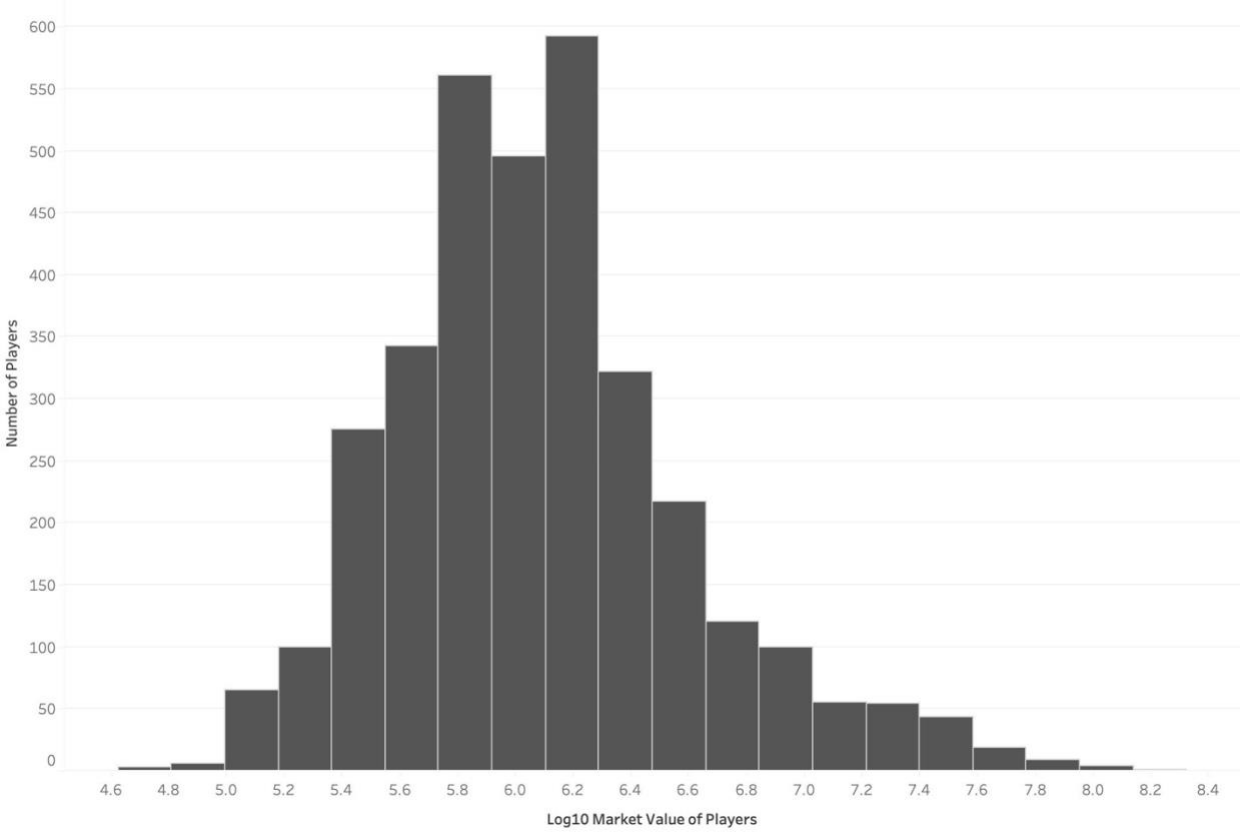
*Distribution of Striker Market Value in Euros*




Initial exploratory data analysis showed that the market value of players varied significantly, with exponential increases in the market value for top players. The right-skewed distribution of market value data showed a stark deviation from the expected normality, which could affect the results of the model. Al-Asadi and Tasdemir (2022) solved this issue by applying a logarithmic transformation to the market value of all players. The new variable (*value_log10*) displayed a normal distribution as seen below and it is now used as the new dependent variable for the multiple linear regression.

**Figure 2**

*Distribution of Striker Market Value after log10 transformation*




The dataset was then checked for missing values in any of the variables. 13 players were found to have missing values for their market value. This was determined to be due the players' status as free agents with no current contracts. As this is a small proportion of players when compared to the population size (n=3,385), these players were disregarded and removed from the analysis. This was done as using mean-value correction would inaccurately reflect against their performance attributes.

Analysis of outliers across the dataset showed the existence of significant natural outliers across all attributes of the dataset. As these outliers were not the result of data entry, or processing errors, and represented true natural variations in the population, they were left as is in the dataset.

## Analysis and Results

Before conducting the multiple linear regression, an exploratory data analysis of the dependent and independent variables was performed to understand and gather some insight on the data. The results of these descriptive statistics are shown in the table below.

**Table 1**

*Descriptive Statistics of Dependent and Independent Variables*

| Variables | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| *value_eur* | 3385 | 45000 | 194000000 | 3111666.174 | 8543809.100 |
| *value_log10* | 3385 | 4.653 | 8.287 | 6.087 | 0.510 |
| *age* | 3385 | 16 | 39 | 25.230 | 4.712 |
| *international_reputation* | 3385 | 1 | 5 | 1.110 | 0.403 |
| *shooting* | 3385 | 40 | 94 | 64.965 | 7.567 |
| *dribbling* | 3385 | 40 | 95 | 65.238 | 7.317 |
| *physic* | 3385 | 33 | 89 | 64.531 | 9.618 |

The initial independent variable *value_eur* (*M* = 3111666.174, *SD* = 8543809.103) is seen to be highly skewed to the right, as the standard deviation is vastly greater than the mean. The maximum value of this variable is also exponentially greater than the mean.

The new independent variable *value_log10* (*M* = 6.087, *SD* = 0.510) that was derived by applying a logarithmic transformation on the *value_eur* variable is observed to have a more normal distribution as the mean is vastly greater than the standard deviation. The minimum and maximum values of *value_log10* are also closer to the mean after the transformation.

Looking at the descriptive statistics of the variable *international_reputation* (*M* = 1.110, *SD* = 0.403), we can see that the value for most of the players in the dataset is closer to 1, with very few players being internationally reputed. This makes logical sense as the vast majority of players do not play at an international level.

The variable *age* (*M* = 25.230, *SD* = 4.712) shows that the average player is in their mid-20s, with some players playing professionally in their teens and other players extending their careers into the late-30s.

The variables *shooting* (*M* = 64.965, *SD* = 7.567), *dribbling* (*M* = 65.238, *SD* = 7.317), and *physic* (*M* = 64.531, *SD* = 9.618) all similarly describe the performance attributes of an average striker. The relatively small standard deviation exhibits the existence of truly exceptional players who can be considered as natural outliers within the striker population.

**Table 2**

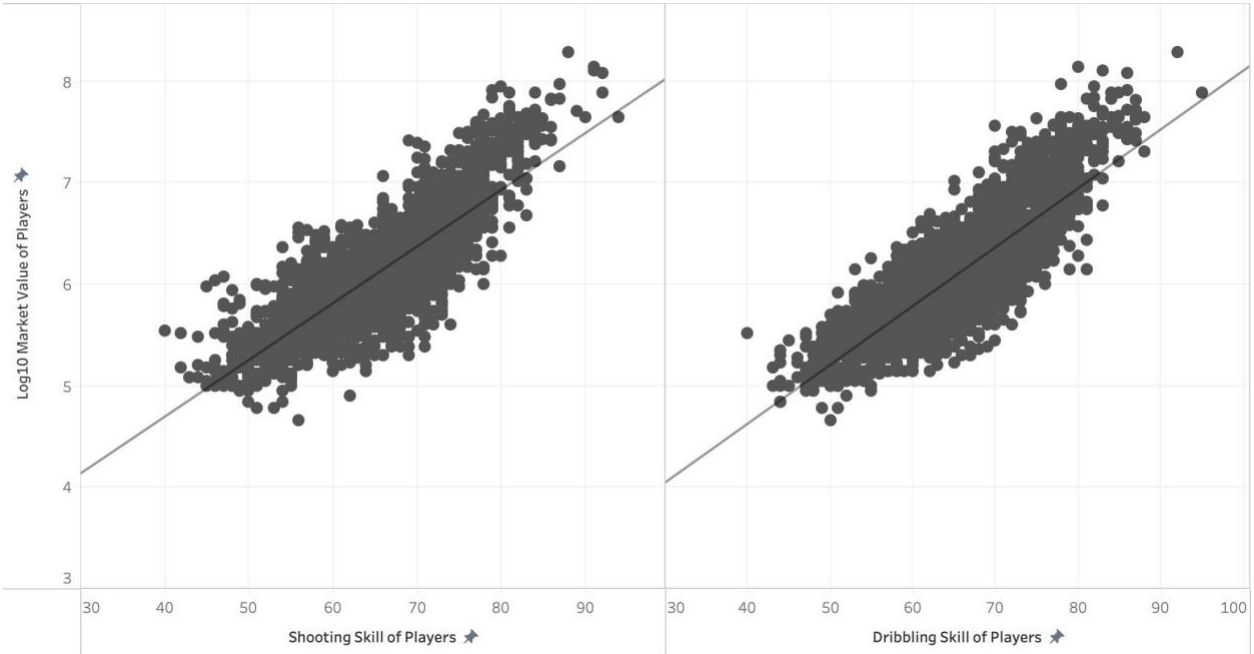
*Correlations of Independent Variables with Dependent Variable value_log10*

| | | *value _log10* | *age* | *international _reputation* | *shooting* | *dribbling* | *physic* |
|---|---|---|---|---|---|---|---|
| *value _log10* | Pearson Correlation | 1 | 0.113 | 0.476 | 0.829 | 0.831 | 0.430 |
| | Sig. (2-tailed) | | <0.001 | <0.001 | 0.000 | 0.000 | <0.001 |

As seen from the above table, the variable that correlates the most with the dependent variable *value_log10* is the independent variable *dribbling*, which is closely followed by *shooting*. The variables *international_reputation* and *physic* exhibit average correlation with *value_log10*, and *age* correlates poorly against the dependent variable.

**Figure 3**

*Scatterplot of Shooting Skill and Dribbling Skill against the log10 Market Value of Strikers*



From the results of the multiple linear regression, it is seen that the chosen combination of independent variables can explain 92.1% of the variation in the market value of the strikers in the population (*R-squared* = 0.921, $F = 7842.358$, $p = 0.000$). The high value of $F$ and the extremely low value of $p$ ($p<0.05$) shows that our model performs significantly better than a model with no predictor variables.

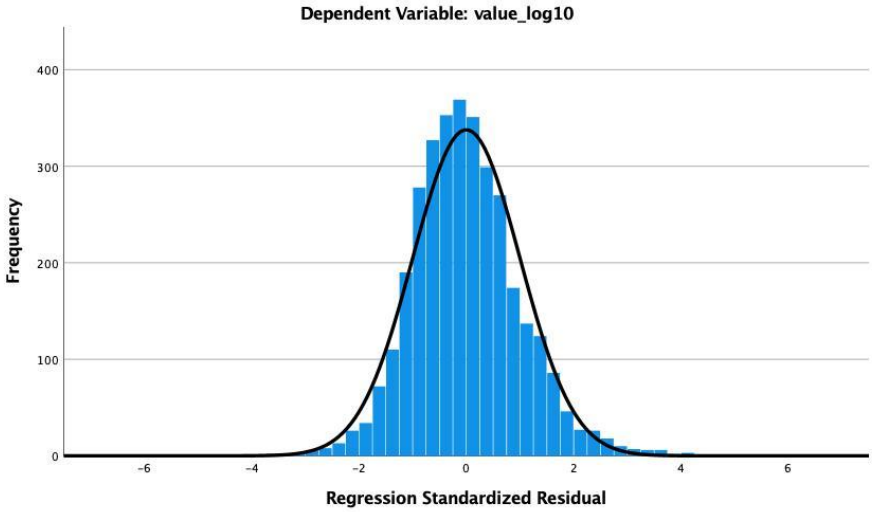
All the independent variables were found to be significant at predicting the *value_log10* variable ($p < 0.001$) and the independent variables *shooting* ($\beta = 0.530$), and *dribbling* ($\beta = 0.477$) were significant individual predictors of the logarithmic market value. Surprisingly, the independent variable *age* ($\beta = -0.424$) significantly described the dependent variable *value_log10* even though it had a poor correlation ($r = 0.113$).

Additional tests were performed to analyze the reliability of the multiple linear regression model based on the assumptions that need to be met. The independence of observations was achieved with a Durbin-Watson score of 1.602 which is within the acceptable range (1.5-2.5). There was no observed multicollinearity between the independent variables as all the individual VIF values (1.294-3.550) were below the safe limit of 5.

The graphs below exhibit that the remaining assumptions for multiple linear regression are met as the residuals (error terms) are normally distributed, linearly related to the predicted values in the P-P plot, and randomly scattered with no observable pattern, the latter confirming homoscedasticity.

**Figure 4**

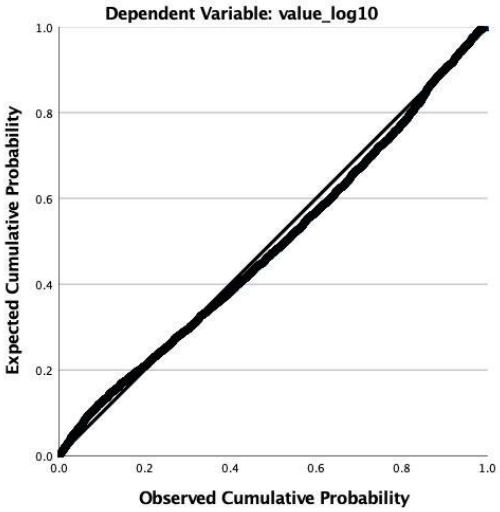
*Frequency Distribution of Regression Standardized Residual*




*Note.* Residuals are seen to be approximately normally distributed.

**Figure 5**

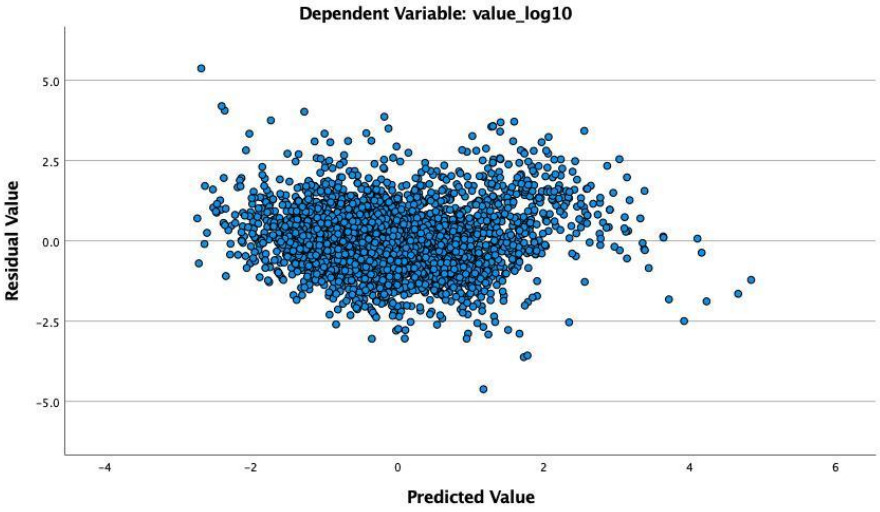
*Normal P-P Plot of Regression Standardized Residual*




*Note.* Linear relationship between Residuals and Predicted values is observed

**Figure 6**

*Scatterplot of Predicted values against Residuals*




*Note.* Random scattering with no obvious pattern confirms heteroscedasticity

## Discussion and Conclusion

Al-Asadi and Tasdemir (2022) conducted a similar study using data from FIFA 20 for a multiple linear regression (*R-squared* = 0.56). In their study, they performed the analysis for players across all positions. This led to the use of common dependent variables for all player positions which could explain the lower accuracy of the model as player market valuation greatly varies across different positions. Their further use of non-linear models such as random forest regression resulted in a far superior predicting accuracy (*R-squared* = 0.95) across all player positions.

Muller et al. (2017) used crowdsourced data from the website transfermarket.com, and introduced a multi-level regression method for evaluating a player's market value across all playing positions. Their model performed well for the lower 90% of transfers, with significant inaccuracies for the top 10%. The authors explained this as an inherent limitation of the data that they collected, which did not capture all the nuances of a player's attributes.

Behravan and Razavi (2021) used the FIFA 20 dataset and performed APSO clustering for all players, to obtain the most relevant features for each position group. Their model was based on the PSO-SVR analysis (*R-squared* = 0.74) with 32 features for predicting the market value of strikers. Further pruning of attributes and improved feature selection would most likely improve the performance of their model.

Accurately estimating a player's market value is an essential task in the process of club transfers. The results from the analysis performed in this report show that careful selection of relevant player attributes can objectively predict the market value of strikers.

Different sets of independent variables can be identified and incorporated for predicting the market value of players in other positions. This process can be further developed for significant and objective estimation of player market values across all possible positions and roles.

The results illustrate the viability of using datasets from football simulators such as EA Sports FIFA to make market value estimation more cost effective for smaller clubs with limited budgets and resources. Data-driven estimations of market value can be effectively used as a baseline in real-life negotiations between clubs and agents.

The model can be improved by incorporating better feature selection, and the use of more complex non-linear methods for prediction. The incorporation of this methodology in real-life transfers could greatly help clubs with recruiting the best talent that meet their requirements while staying within the limitations of Financial Fair Play.

## Personal Reflection

The use of real-life datasets from various sporting business requirements across different and exciting fields throughout the course content of the module inspired me to identify a research question that can provide results for real world applications in sport business.

The most challenging aspects of the module were to quickly understand the application of statistics in sport along with the nuances of analytical methods and processes. The introduction to analytical software such as SPSS and Tableau greatly helped me effectively understand the concepts and methodologies in this module at a much faster rate.

If given this assignment again, I would approach the selection of (the mostly unused) independent variables differently with an aim to maximize the synergy between attributes in a way that can comprehensively explain the variances in the dependent variable with greater accuracy.

## References

Al-Asadi, M. A., & Tasdemir, S. (2022). Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*, *10*.

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing - a Fusion of Foundations, Methodologies and Applications*, *25*(3), 2499–2511.

Carmichael, F., & Thomas, D. (1993). Bargaining in the transfer market: theory and evidence. *Applied Economics*, *25*(12), 1467–1476.

Cotta, L. (2016). Using FIFA soccer video game data for soccer analytics. *Proc. Workshop Large Scale Sports Anal.*, 1–4.

Gerrard, B. (2001). A new approach to measuring player and team quality in professional team sports. *European Sport Management Quarterly*, *1*(3), 219–234.

Miller, T. W. (2015). *Sports Analytics and Data Science: Winning the Game with Methods and Models*. Pearson.

Muller, O., Simons, A., & Weinmann, M. (2014). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research*, *263*(2), 611–624.

Singh, P., & Lamba, P. S. (2019). Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. *Publication Cover  Journal of Discrete Mathematical Sciences and Cryptography*, *22*(2), 113–126.

Skinner, J., Edwards, A., & Corbett, B. (2014). *Research Methods for Sport Management* (1st ed.). Routledge.

Thrane, C. (2019). *Applied Regression Analysis: Doing, Interpreting and Reporting* (1st ed.). Routledge.

Tosato, M., & Wu, J. (2018). An application of PART to the Football Manager data for players clusters analyses to inform club team formation. *Big Data and Information Analytics*, *3*(1), 43–54.