# Predictive Modelling on the Market Value of Soccer Players

##### **Anthony Shenouda**

Computing & Software 764, McMaster University

December 15th, 2023

https://github.com/AnthonyShenoud/MarketValueSoccerPlayers

***Abstract*-** The realm of soccer player valuation constitutes a landscape characterized by enormous spending, fueled by the pursuit of top talent. However, organizations are cursed by the problem of biased player reputations, impeding accurate assessments of a player’s worth. This study navigates this intricate terrain by analyzing the relationship between individual attributes, particularly technical skills, and a player's market value. Addressing the challenge of bias, this research dissects the specific skills and statistics that influence a player's worth, while considering the impact of physical, mental, and technical attributes. To form the groundwork, data cleaning techniques sourced from the official FIFA website are employed including currency alignment, duplicate detection, and domain knowledge. Next, correlation analyses unveil the most pivotal attributes, followed by outlier detection mechanisms ensuring data integrity. Finally, with regression methodologies and Python's scikit-learn, the study culminates in an evaluative model that predicts player valuation, emphasizing technical skills as prime indicators of worth. The results of this study detail an accurate estimation of player values based solely on technical prowess, offering insights into market efficiency, risk assessment precision, and data-driven player contract negotiations. This study offers a glimpse into the transformative potential of skill-centric valuations in sports, highlighting their implications in revolutionizing the industry.

I

1. Introduction

n the world of European soccer, clubs allocate vast sums of money and engage in exhaustive scouting processes to secure top-tier players. There is a need for an accurate estimation of a player's market value to avoid penalties and stay within budget, but this estimation continues to be an enigma for clubs. The staggering investments and efforts highlight the critical necessity for precise player evaluations, driving the purpose of this study. The goal is to unravel the interplay of individual attributes, ranging from technical skills, and physical capabilities, to the incalculable attributes of mentality—that dictate a player's worth in the market.

The central issue this research aims to resolve is the large biases within player reputations that often skew assessments, compounded by the elusive nature of isolating specific skills or statistics dictating a player's value. Another challenge stemming from these issues is the weighting process of a player’s attributes. This study seeks to eliminate the layers of complexity associated with accurate player assessments by ensuring there is equity and differentiation between the attributes being analyzed. Ensuring that each attribute will be weighted equitably presents a new challenge this study will address—which attributes should be weighted more heavily in the player valuation process.

The outcomes of this study extend far beyond the parameters of soccer economics, offering transformative potential for player evaluation methodologies. Impacts include the refinement of transfer market efficiency and risk assessment precision as well as the use of data-driven negotiations for player contracts. These insights are both relevant and applicable in broader domains such as fantasy sports and betting, reshaping strategies, and predictive analytics.

The proposed solution begins with a methodical and structured pipeline, commencing with parsing procedures that extract data from the FIFA website with thousands of players and their 50+ attributes. The data is then pulled into a Python environment where data cleaning procedures begin. This phase encompasses aligning currencies, detecting, and rectifying duplicates, and the removal of attributes based on domain knowledge and relevance. The next step entails employing correlation measures to identify the most impactful attributes—specifically, isolating the top 6 attributes that span across mental acuity, physicality, and technical finesse.

The process undergoes further refinement through a comprehensive comparative analysis of outlier detection methods, comparing the Interquartile Range (IQR) and Percentile with Winsorization approaches to ensure data integrity. The study also ventures into assessing various regression techniques, analyzing both the normal and logarithmic data values to discern patterns and trends. Leveraging the capabilities in Python's scikit-learn library, the research culminates in the construction of an evaluation measure. Here linear regression is compared against the predictive technique of random forest regression.

This comprehensive approach aims to provide a solution to the immediate challenge of player valuation and serves as a foundational blueprint, demonstrating the integration of data management practices into the sports domain. By utilizing the predicting model to estimate a player's value based solely on technical characteristics, this research not only illustrates the transformative potential within soccer player valuations but also sets the stage for the integration of more advanced techniques in future research.

1. Related Works
   1. Introduction of Ulaş Özen's Work:

Ulaş Özen, a recent graduate student from Istanbul University, laid the groundwork for predicting soccer player market values. His model utilized parsing techniques and attribute selection aimed at forecasting a player’s worth. Özen's work initially started on player valuation methodologies, but certain limitations persisted within his approach.

* 1. Limitations

Building upon Özen's foundation, this research addresses several key limitations to enhance the prediction of soccer player market values. Özen's parsing technique has become outdated over time, prompting the need to incorporate modern parsing methods in data extraction. Additionally, Özen's attribute weighting didn't align accurately with player skills; this model implements sophisticated calculations and robust modelling to discern and select the six attributes crucial in defining a player's skills.

* 1. Expanded Methodologies

Expanding beyond Özen's framework, this model incorporates two distinct outlier detection techniques: the Interquartile Range (IQR) and Percentile methods. Moreover, it integrates two regression methodologies, including analysis on both normal and logarithmic data values, and employs Python's scikit-learn to introduce machine learning techniques. Both techniques significantly refine accuracy and mitigate biases in player valuation. This paper also offers a study comparing multiple data management techniques within sports analytics allowing readers to educate themselves on how various methodologies can be applied.

* 1. Divergent goals

Özen included attributes such as player popularity to formulate his algorithm, which skews the results of a player’s value when using his model. This model, however, takes on a different approach. By removing inherent biases, such as excluding attributes like International Reputation and selecting non-overlapping attributes, it calculates what a player's actual value should be. This shift eliminates overrepresented attributes, ensuring a more accurate and unbiased estimation of a player's worth.

* 1. motivation for solution

The limitations and gaps identified in Özen's model serve as the primary motivation for the innovations and refinements of this study. By rectifying parsing techniques, recalibrating attribute selection, eliminating biases using data management techniques, and offering a comparative study on various methodologies, this paper aims to set new standards for precise, bias-free player valuations within the domain of soccer economics.

1. Proposed Solution
   1. The Dataset

The dataset (Figure 1) utilized for this study serves as a comprehensive repository sourced directly from the official FIFA website. It encompasses a vast array of over 50 player attributes ranging from traits such as shot power, passing, and dribbling that define a player's overall value. This dataset provides insight into the skill sets and characteristics defining modern soccer players. Collecting data from the most recent 2023-2024 season across the top five European leagues, this dataset encapsulates the profiles of nearly 3000 players, offering a rich pool of talent from the world’s most elite soccer competitions. Notably, the dataset comprises a blend of both synthetic and real-world data, combining meticulously curated statistics with real-time performance metrics. This data bridges the gap between theoretical player attributes and on-field realities, providing an understanding of player skills and values.

A screenshot of a computer

Description automatically generated

Figure 1 FIFA Dataset

* 1. Parsing player data

The procedure for acquiring, parsing, and processing the dataset starts with direct extraction from the official FIFA website, leveraging Python's Beautiful Soup web-scraping library. Post-extraction, that data is structured and organized in an Excel file, a meticulously organized repository meticulously organized, facilitating seamless access and manipulation for subsequent analyses. The combination of web-scraping techniques and Excel storage ensured no values were lost. However, issues with duplication, currency conversions, and improperly formatted values were next to tackle.

* 1. DATA CLEANING

The data cleaning phase aims at refining the acquired dataset to ensure its accuracy and suitability for analysis. The process begins by employing Pandas' duplicate function which detects and eliminates any redundant entries. Next, a critical aspect addressed was the uniform representation of player values, which often utilized ‘M’ and ‘K’ to denote millions and thousands and ‘€’ for euros. Therefore, a function was developed to standardize these values into a complete numeric form. The next issue was with the attribute values. Throughout the season, players would improve or underperform and their characteristics would be updated accordingly. This was represented by the current weighting followed by a ‘+’ or ‘-’ number representing the change (i.e., 84+1 or 76-3). Another function was developed to rectify this issue. Furthermore, non-numeric columns such as player positions were removed. While these columns held relevance in player profiling, their exclusion was necessary for calculations of correlations as they could not be quantified.

* 1. Correlation Measure

This part of the process involved calculating the correlation matrix between the player value and each attribute in the dataset. This process unveiled the relationships between a player's market worth and their skills. It provided insights into which attributes strongly influence a player's overall value, guiding more focused player assessments.

A diagram of heatmap

Description automatically generated

Figure 2 Correlation Matrix Heat Map using Market Values

A screenshot of a computer program

Description automatically generated

Figure 3 Correlation Matrix Results

* 1. Selecting attributes

The process of selecting attributes led to the identification of six key attributes—'Pace / Diving', 'Shooting / Handling', 'Passing / Kicking', 'Dribbling / Reflexes', 'Defending / Pace', and 'Physical / Positioning'. These attributes were chosen due to their encapsulation of multiple underlying attributes within FIFA's structure, ensuring an accurate and comprehensive representation of player skills. For instance, 'Shooting / Handling' incorporates a blend of attributes like volleys, penalties, long shots, shot power, finishing, and attack positioning, each carrying a specific weightage that aligns with FIFA's weighting system (Figure 4). Similarly, each selected attribute header adeptly encapsulates a cluster of related attributes, effectively capturing all skill sets weighted accurately. This is one of the key differences between this study and Özen's.

A screen shot of a player attributes

Description automatically generated

Figure 4 Attributes Selected

* 1. Outlier Detection Algorithms

For data integrity, two outlier detection techniques were employed: the Interquartile Range (IQR) method and percentile with winsorization. The IQR method focuses on identifying outliers by establishing the range between the first and third quartiles of a dataset, allowing for the detection of values lying beyond a specific range [Goyal, Chirag]. On the other hand, the Percentile method, coupled with Winsorization, sets thresholds based on percentile values, replacing outlier values with predefined boundary values. This effectively minimizes their impact on the analyses [Goyal, Chirag].

* 1. Regression Techniques

The section employed two distinct regression techniques: Ordinary least squares (OLS) regression and logarithmic regression, each offering unique advantages in modelling player valuation. OLS regression, a classic statistical method, establishes the linear relationship between a player's value and the selected attributes. It optimizes coefficients to minimize the sum of squared differences between observed and predicted values [Vasudev, Rakshith]. Conversely, logarithmic regression operates by transforming the data, fitting a curve that better captures potential non-linear relationships between player attributes and valuation [Vasudev, Rakshith]. This approach accommodates non-normally distributed data offering in this situation a more flexible and accurate model.

* 1. Machine Learning Methodologies

The final step was to incorporate machine learning methodologies from the Python scikit-learn library. This analysis used linear regression and random forest regression to create the player valuation model. Linear regression, a fundamental yet powerful technique, aims to establish a linear relationship between a player's value and the selected attributes. By optimizing coefficients, it constructs a linear equation to predict a player's worth based on attribute values [Varghese, Danny]. Conversely, random forest regression operates as an ensemble learning method, leveraging the strength of multiple decision trees to predict player values [Varghese, Danny]. It uses the collective insights of numerous trees; each trained on different subsets of the data and aggregates their predictions for an accurate estimation [Varghese, Danny].

1. Experiments
   1. comparison of regression techniques

The R2 score measures how well the regression model explains the variability of the dependent variable based on the independent variables. It ranges from 0 to 1, with higher values indicating a better fit for the model. As explained in Section 3 Proposed Solution, Regression Techniques, Logarithmic Regression outperforms the benefits of OLS with non-normally distributed data.

A screenshot of a computer

Description automatically generated

Figure 5 OLS vs Logarithmic Regression with IQR

* 1. Model results before machine learning and Özen's

The table below highlights the benefits of this study. It shows how each incremental step of the pipeline aids in improving the accuracy of the model from data cleaning, and outlier detection, to logarithmic regression. Özen's r2 score up to this point was 0.696. This would make sense as he opted to use attributes more strongly correlated to market value such as the player's international reputation score. However, considering that bias was removed from this model, the results till here are strong.

A screenshot of a computer

Description automatically generated

Figure 6 R Scores of Model Pipeline

* 1. Comparison of machine learning techniques and outlier detection algorithms

Using the logarithmic data from the previous step; the table below outlines the advantages of random forest regression. With a r score goal of 1, random forest does an excellent job with this model. This experiment also compared Özen's results using IQR with this model's percentile outlier detection. His linear regression machine learning model had a score of 0.696 compared to this model's 0.625 (shown above and below) however, this model had a smaller mean square error. This is due to the benefits of the percentile outlier detection algorithm over the IQR algorithm as discussed in Section 3 Proposed Solutions, Outlier Detection Algorithms.

A screenshot of a computer

Description automatically generated

Figure 7 Random Forest vs Linear Regression with Percentile Outlier Detection

1. Applications

The comprehensive advancements and refinements integrated into this model for predicting the market value of soccer players extend beyond player valuation. This model is applicable to soccer economics and more.

* 1. Transfer Market Efficiency:

The refined accuracy in player valuation can significantly enhance the efficiency of transfer markets. Clubs, armed with more precise player valuations devoid of biases, can make more informed decisions when acquiring or selling players. This can streamline transfer negotiations and facilitate smoother transactions, optimizing resource allocation and investment strategies.

* 1. Risk Assessment and Strategic Planning:

Accurate player valuations are critical for risk assessment and strategic planning within soccer clubs. Understanding a player's true worth aids in risk mitigation, enabling clubs to navigate negotiations and investment decisions with a heightened sense of confidence and clarity.

* 1. Data-Driven Player Contract Negotiations:

The model's emphasis on unbiased player valuations empowers data-driven negotiations for player contracts. Clubs and players can leverage these precise valuations to formulate more equitable and strategic contract offers, ensuring a fair and just compensation structure aligned with a player's real value.

* 1. Fantasy Sports and Betting:

The ramifications extend into the realm of fantasy sports and betting, revolutionizing strategies, and predictions. Enhanced accuracy in predicting player values offers enthusiasts a framework for constructing fantasy teams and making informed betting decisions based on unbiased player valuations.

* 1. Future Advancements in Soccer Analytics:

Beyond immediate applications, this paper paves the way for future advancements in soccer analytics. The emphasis on eliminating biases and comparison of methodologies sets a precedent for further research endeavours aiming to refine and innovate player valuation methodologies.

* 1. Wider Applicability in Sports Economics:

The principles and techniques embedded within this model hold potential applications beyond soccer economics. The refined techniques in data management, unbiased valuation, and analytics can be adapted and applied to evaluate player values in other sports domains, contributing to a broader understanding of sports economics.

1. Conclusions

In summary, the advancements and refinements integrated into this model for predicting the market value of soccer players represent significant progress in the domain of player valuation methodologies. Through an analysis and comparison of existing methodologies, several key lessons and avenues for future work have emerged.

​The foundational work by Ulaş Özen's and the enhancements introduced in this model highlights the growth of player valuation. Key takeaways include—knowledge of website parsing techniques, the integration of domain knowledge with data science and data science with machine learning, and other data cleaning and mining techniques. This model used as a study demonstrates the importance of data cleaning techniques including duplicate detection and data consistency. It also outlines the pros and cons of other formal techniques like IQR and Percentile Outlier Detection, linear vs logarithmic regression, and linear and random forest machine learning algorithms.

* 1. Future work:
     1. Integration of Advanced Statistical Techniques:

Future work includes integrating other advanced statistical techniques and machine learning algorithms. Exploring methodologies like neural networks or ensemble learning could potentially enhance the accuracy and predictive power of player valuations. Aside from an increase in precision, more research into how computing and software techniques apply to real-world problems converts this domain from student research to a million-dollar project.

* + 1. Incorporating Dynamic Factors and Contextual Data:

Incorporating dynamic factors such as a player's performance trajectory over time could add layers of depth to player valuations. This can be achieved by training the model on an individual's change in skills over time utilizing historical data from the official FIFA page. Future models could strive to integrate these aspects for a more comprehensive estimation of player worth.

* + 1. Expansion to Other Sporting Domains and Industries:

The model's principles and methodologies go beyond soccer economics, offering potential applications in other sports and even industries beyond sports. Future research could explore adapting and applying these techniques to evaluate player values in diverse sporting arenas or extend the model's framework into domains like talent evaluation in corporate settings.

In conclusion, this paper represents a pivotal step in revolutionizing player valuation methodologies within soccer economics. The research shows the importance of modernized techniques, the necessity for bias-free valuations, and the vast potential for further advancements and applications. Embracing advanced statistical methodologies, incorporating dynamic factors, and extending applications beyond soccer present promising trajectories for future research in refining player valuation methodologies and broader applications.

References

1. "Chatgpt - Openai." ChatGPT, chat.openai.com/auth/login. Accessed 7 Dec. 2023.
2. FIFPlay. "FIFA 22 Attributes Explained." FIFPlay, www.fifplay.com/fifa-22-player-attributes/.

Accessed 15 Dec. 2023.

1. Goyal, Chirag. "Outlier Detection & Removal: How to Detect & Remove Outliers (Updated 2023)." Analytics Vidhya, 14 Aug. 2023, [www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-](http://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-) python- code/#h-percentile-method
2. "Players."SoFIFA,sofifa.com/players?type=all&lg%5B0%5D=13&1g∞5B1%5D=16&1g%5B2%5D=19&1g%5B3%5D=31 &lg95B4%5D=53&col=vl&sort=desc&showCol%5B%5D=phy. Accessed 11 Nov. 2023.
3. Varghese, Danny. "Comparative Study on Classic Machine Learning Algorithms." Medium, Towards Data Science, 10 May 2019, towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9fi6ab222.
4. Vasudev, Rakshith. "How Are Logistic Regression & Ordinary Least Squares Regression (Linear Regression) Related?" Medium, Towards Data Science, 5 June 2018, towardsdatascience.com/how-are-logistic-regression-ordinary-least- squares-regressionrelated-Ideab32d79f5.
5. Özen, Ulas. "Predicting Football Players' Market Value Using Machine Learning." Medium, Medium, 3 June 2023, yulasozen.medium.com/predicting-football-players-market-value-using-machine-learning-b28be29891e.

Authors

**First Author** – Anthony Shenouda, MASc., Software Engineering, McMaster University, shenoa2@mcmaster.ca.

**Correspondence Author** – Dr. Fei Chiang, Associate Professor, McMaster University, fchiang@mcmaster.ca,