Predicting Player Market Value Based on Performance Metrics

Felipe Villegas

***Abstract*— This project investigates whether structured performance metrics can be used to predict season-over-season market value changes for professional soccer players. Using data collected from public sources, an end-to-end machine learning pipeline was built covering data extraction, exploratory analysis, feature engineering, and predictive modeling. Multiple regression-based and non-linear models were evaluated, including random forests, support vector machines, and neural networks. Results show that model performance varies significantly by player position, with stronger predictive power for forwards than for midfielders and defenders. Feature importance analysis highlights past market value and attacking contributions as key drivers, while also revealing the limitations of purely quantitative performance data for player valuation. The findings emphasize both the potential and constraints of applying machine learning to real-world sports valuation problems.**

1. PROBLEM STATEMENT

The soccer transfer market, a multi-billion-dollar global industry, revolves around estimating player market values. Market value serves as a critical metric for clubs, agents, and stakeholders to evaluate players, negotiate contracts, and structure trades. Understanding the underlying factors that influence changes in market value over time has significant implications for optimizing transfer strategies and scouting efforts.

This study aims to test the hypothesis that a soccer player's market value change over one season in their local league is directly influenced by their performance metrics during that period. By examining league-specific data such as goals, assists, defensive contributions, and other performance indicators, we attempt to quantify this relationship using machine learning models. This project investigates whether these numerical metrics alone are sufficient to predict market value changes or whether external qualitative and contextual factors dominate.

While the results of this study provide insights into the role of quantitative performance metrics in market value prediction, the limitations of excluding qualitative and contextual variables, such as international tournament performance or media presence, highlight gaps in current methodologies.

2. DATA SOURCE

The data used in this study was obtained from FBRef and soFIFA, online platforms recognized for its comprehensive repositories of soccer player statistics. Using automated scraping tools built with Selenium, data was collected from two time points: the present - end of the 2023/2024 season and the past - start of the 2023/2024 season. The present data contained player profiles, including market values, demographic characteristics, and performance metrics, while the past data provided historical market values for comparison and analysis of changes over time. The features that were chosen for the project evaluation were as follows:

Player	Name of player
Team	Team where player performs
Comp	League where team plays
Age	Age of player
Nation	Nationality
Pos	Most popular position on the field
MP	Matches Played
Starts	Games started by player
Min	Minutes played
Gls	Goals Scored
Ast	Assists
CrdY	Yellow Cards
CrdR	Red Cards
PrgC	Progressive carries
PrgP	Progressive passes

PrgR	Progressive Passes Received
Sh	Shots total (no PK)
G/Sh	Goals per Shot
Cmp%	Pass completion percentage
PrgDist	Progressive passing distance
SCA	Shot-creating actions
GCA	Goal-creating actions
TklW	Tackles Won
Blocks	Blocking ball by standing in its path
Tkl+Int	Players tackled + interceptions
Succ%	Percentage of successful take-ons
Reputation	International player reputation
Ball_control	Qualitative ball control metric
market_value_present	Market value end of season
market_value_past	Market value start of season

The data collection process faced significant challenges. CAPTCHA interruptions were frequent, requiring manual intervention during scraping. URL redirection issues arose when profiles for specific timepoints did not exist, often leading to the redirection to an unintended current profile. To address these issues, random delays and fake desktop user agents were implemented to mimic human behavior and minimize detection. Additionally, layouts for mobile and desktop user agents differed, which led to inconsistencies in data collection; this was solved by limiting user agents to desktop-only configurations.

The final dataset contained over 2,000 players from the 5-top European league records with detailed performance statistics, demographic details, and market value changes. Data preprocessing steps ensured the dataset was ready for machine learning, including handling missing values, encoding categorical variables, and normalizing performance metrics for consistency.

3. METHODOLOGY

To explore the hypothesis that market value change is influenced by performance metrics, the study employed a combination of exploratory data analysis (EDA), feature engineering, and machine learning techniques.

The first stage of the project involved EDA to understand the dataset's structure and uncover trends and patterns. Boxplots were generated to compare performance metrics across different player positions and features, revealing clear differences in statistical distributions.

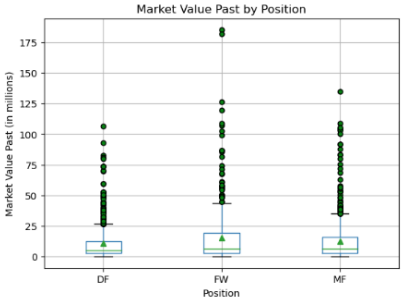

Figure 1. Market Value by Position

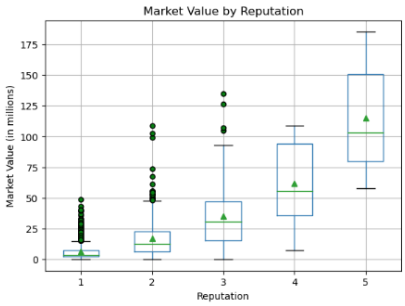

Figure 2. Market Value by International Reputation

Forwards exhibited higher contributions in metrics such as goals and assists, while defenders excelled in tackles and blocks. Correlation heatmaps provided insights into relationships between metrics and market value change, showing stronger correlations for forwards, particularly between goals, assists, and past market value.

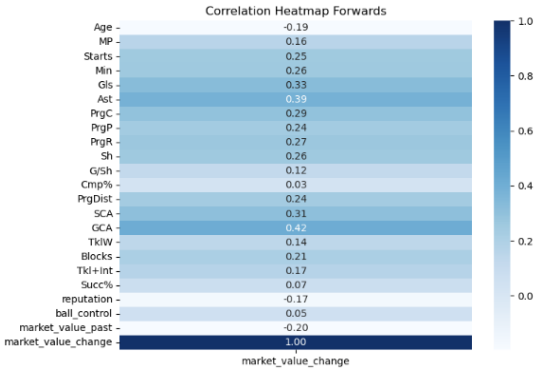

Figure 3. Correlation Forwards

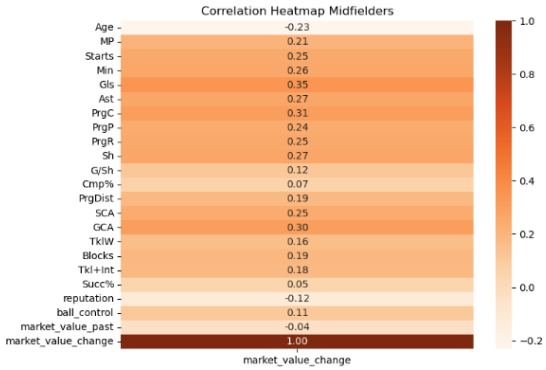

Figure 4. Correlation Midfielders

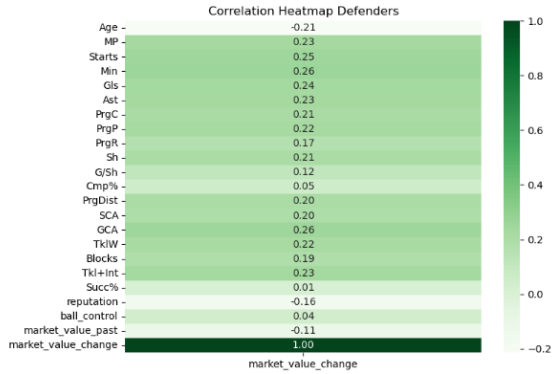

Figure 5. Correlation Defenders

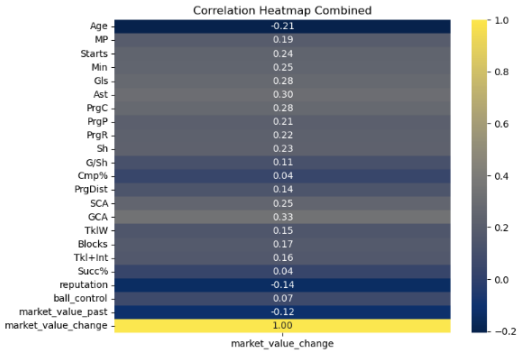

Figure 6. Correlation Combined

Furthermore, radar charts were used to visualize the distinct patterns of performance metrics by position, emphasizing the specialized nature of each role on the field.

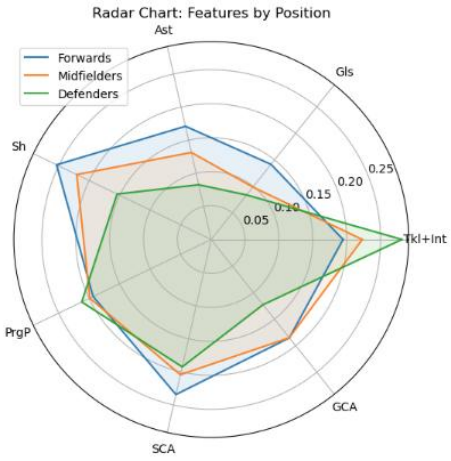

Figure 7. Dominant Features by Position

In the second stage, feature engineering transformed the raw dataset into a machine-learning-ready format. Categorical variables, such as league, were one-hot encoded, ensuring minimal multicollinearity. Numerical columns, including ball control and passing accuracy, were handled where missing values occurred.

Machine learning models were then employed to evaluate the hypothesis. The modeling process began with linear regression, which serves as a baseline due to its interpretability and grounding in statistical learning theory. However, the assumption of a linear relationship between variables proved insufficient for the non-linear dependencies in soccer performance data, necessitating the use of more complex models.

To address these limitations, random forests, support vector machines (SVM), and neural networks were employed. Random forests, an ensemble learning method, build multiple decision trees to capture non-linear interactions while reducing overfitting through averaging. Hyperparameter optimization for random forests was conducted using GridSearchCV, tuning parameters such as tree depth, the number of trees, and minimum split size. SVMs, applied kernel transformations to capture non-linear patterns but struggled with sensitivity to data scaling and hyperparameter tuning. Neural networks, with their ability to approximate high-dimensional relationships through interconnected layers, offered flexibility but required careful adjustments to avoid overfitting given the limited size of the dataset.

Models were trained using an 80/20 split of training and testing data. Additionally, position-specific models were developed for forwards, midfielders, and defenders to account for the unique nature of performance metrics for each role. The hypothesis was that dividing the dataset by position would improve prediction accuracy, as certain metrics (e.g., goals and assists) are more relevant for forwards, while others (e.g., tackles and blocks) are critical for defenders. Feature importance rankings derived from the random were further used as attempts to simplify the model by using only the most predictive features, such as past market value, goals, and assists, to minimize noise and reduce overfitting.

4. EVALUATION AND FINAL RESULTS

The evaluation of the models highlighted the complexity of predicting market value changes solely based on league performance metrics. Despite rigorous preprocessing, feature selection, and model tuning, the testing set results revealed relatively low R^2 values (0.3–0.4) for most models, indicating limited predictive power. Random forests consistently outperformed other models, achieving the highest R^2 values across both the overall dataset and position-specific datasets. However, these values were lower than anticipated, underscoring the influence of external factors beyond league performance metrics.

The positional analysis provided mixed results. For forwards, the random forest model yielded an improved R^2 value of 0.64, likely due to the strong correlations between measurable metrics such as goals and assists and market value changes. In contrast, models for midfielders and defenders exhibited lower R^2 values (0.00–0.30), reflecting the qualitative nature of their roles. Metrics like positioning, tactical awareness, and creativity, which are critical for midfielders and defenders, are not easily captured in numerical datasets, thereby limiting the models' ability to predict market value changes effectively for these positions.

Further analysis of feature importance revealed that past market value, goals, and assists consistently ranked as the most influential predictors across models. However, these findings varied by position, with defensive actions such as blocks and tackles being more significant for defenders. Visualization techniques, such as radar charts and correlation heatmaps, highlighted these differences, underscoring the distinct performance patterns across positions.

The results suggest that while numerical performance metrics are valuable predictors for specific roles, such as forwards, they fall short of capturing the multifaceted nature of player performance. The low R^2 values in testing reflect the significant impact of external and qualitative factors—such as international

competition, media narratives, and personal circumstances—that were not accounted for in this dataset. This highlights the limitations of quantitative approaches and suggests that future models should integrate contextual and qualitative data to improve predictive accuracy.

```
Results for Full:
Score (R²): 0.35

Top 3 Features:
               Feature  Importance
14                 GCA    0.141143
21  market_value_past    0.138194
0                  Age    0.131070

Results for FW:
Score (R²): 0.64

Top 3 Features:
               Feature  Importance
21  market_value_past    0.211258
14                 GCA    0.174780
5                  Ast    0.077229
```

Figure 8. Results for Full and FW

```
Results for MF:
Score (R²): -0.02

Top 3 Features:
               Feature  Importance
4                  Gls    0.144680
21  market_value_past    0.143983
11                Cmp%    0.129006

Results for DF:
Score (R²): 0.33

Top 3 Features:
               Feature  Importance
21  market_value_past    0.123459
0                  Age    0.114464
3                  Min    0.089305
```

Figure 9. Results for MF and DF

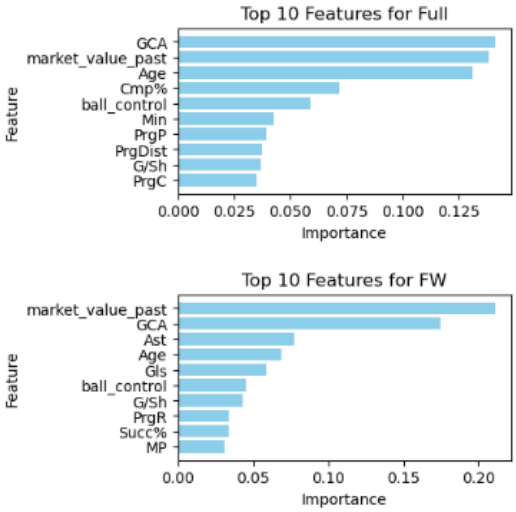

Figure 10. Top Features Full and FW

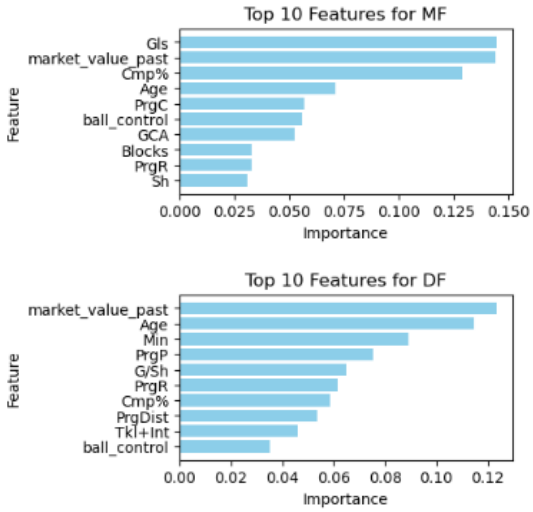

Figure 11. Top Features MF and DF

The results support the reality of numerical performance metrics being valuable predictors for certain positions, such as forwards, but insufficient to capture the multifaceted nature of player performance for others.

5. CONCLUSIONS

This study tested the hypothesis that market value changes in soccer players are influenced by league performance metrics, using machine learning models to quantify these relationships. While the models provided moderate predictive accuracy for forwards, where numerical metrics such as goals and assists are directly linked to valuation, they were less effective for midfielders and defenders. These results show the qualitative and context-dependent nature of certain roles, emphasizing that soccer performance cannot be fully captured by numerical data alone.

The project faced several challenges, particularly in data collection and model evaluation. CAPTCHA interruptions and URL redirections complicated web scraping, while the dataset's limitations—excluding

external factors such as media coverage, international performance, and injuries—restricted the models' predictive power. Attempts to improve accuracy by selecting top features and training position-specific models yielded mixed results, with only marginal improvements for forwards and lower performance for other positions. These findings highlight the complexity of market value prediction and the importance of factors outside the scope of the available data.

Despite these challenges, this project demonstrates the potential of machine learning in sports analytics and lays the base for future research. Expanding datasets to include qualitative and contextual factors, alongside more advanced modeling approaches, could significantly enhance predictive accuracy. Ultimately, this study reinforces the need for a comprehensive approach, combining data-driven insights with the qualitative understanding required to evaluate the nature of soccer player performance and valuation.