Forecasting Soccer Player Value via Integration of Injury History with Machine Learning

Alec Zhang

500 Saratoga Avenue, San Jose, CA, 95129

The Harker School

ABSTRACT **3**
1. INTRODUCTION **4**
2. DATA SELECTION **5**
2.1 Data Extraction 5
2.2. Data Concatenation 6
2.3. Data Cleaning 7
2.3.1 Injury Duration Model 7
2.3.2 Player Value Model 7
3. MODEL SELECTION **8**
3.1 XGBoost model 8
3.2 Model Building 9
3.2.1 Injury Duration Model 9
3.2.2 Player Value Model 11
4. RESULTS **13**
5. DISCUSSION **16**
5.1 Findings 16
5.2 Limitations 17
5.3 Future Research and Implementation 18
6. CONCLUSION **18**

ABSTRACT

The soccer transfer market is highly volatile and influenced by player performance and injuries, so accurately assessing player value is crucial for clubs. While existing machine learning models focus on performance metrics, this study emphasizes the significance of injuries in player valuation. Utilizing Extreme-Gradient Boosting (XGB) algorithms, this paper integrates player statistics and injury data spanning multiple seasons from top European leagues. Through meticulous data extraction, cleaning, and model building, the study aims to provide clubs with a comprehensive evaluation tool for player acquisitions. The proposed model was tested across one to five-year windows. The 4-year sliding window performed the best (R-squared = 0.856, MAE ≈ $3.82 million). The findings indicate that while traditional performance statistics drive most predictive accuracy, current publicly available injury data did not enhance model performance.

1. INTRODUCTION

Soccer is the most popular and widely watched sport in the world, with transfers occurring nearly every day across top leagues. Effectively assessing the value of a soccer player includes gauging the potential risks associated with the player, which poses a considerable challenge for sports teams. In recent years, soccer clubs have made substantial financial investments in player acquisitions, only to find that these signings failed to benefit the team or proved costly. This issue has become increasingly prevalent, particularly with the rise in player value in the 21st century. In the Premier League, the transfer market exhibits characteristics similar to those of economic bubbles, such as the growth in TV broadcasting revenue. These significant changes in player prices force clubs to be even more confident in their purchases [1]. While several machine learning models incorporate statistical metrics such as goals scored, assists provided, and red cards received to predict a player's value [2,3,4], they often overlook a critical factor: injuries. On the other hand, other research has focused solely on injuries [5,6], but none have incorporated injuries and statistics into player valuation. This research aims to bridge the gap between injuries and statistics to accurately predict a player's value.

Injuries play a pivotal role in shaping a soccer player's career trajectory and influencing their future performance. A comprehensive study from 2001 to 2008, encompassing 23 professional soccer teams across Europe, revealed that each player sustained, on average, approximately 2.0 injuries per season. Notably, re-injuries accounted for approximately 12% of these injuries, resulting in extended periods of absence compared to initial injuries [7]. Furthermore, injuries exact a substantial economic toll on English Premier League teams, with an estimated annual loss of approximately £45 million ($57,087,000) [8].

Given the financial losses incurred due to injuries, the primary objective of this paper is to develop a precise and reliable system for predicting the value of a soccer player. An Extreme Gradient Boosting (XGB) model was used to assist professional soccer coaches and scouts in determining whether a player is a good fit for their team. It provides an in-depth player evaluation by factoring in their historical statistics, performance metrics, and injury history. Additionally, a predictive model can identify various risk factors associated with a player, such as their susceptibility to injury, age, and prior performance history. By leveraging such insights, clubs can proactively manage their player assets, make informed decisions regarding player acquisitions, and identify performance trends that may lead to fluctuations in player value in subsequent seasons. This advancement in player evaluation represents a critical stride toward enhancing the efficiency and effectiveness of talent recruitment and management in professional soccer. This multifaceted approach enables clubs to identify undervalued players with significant potential and determine optimal times to sell existing assets. This study aims to systematically evaluate whether public injury history data can improve player valuation, compare sliding-window horizons, and analyze feature importance to identify what the model is learning.

2. DATA SELECTION

This step ensures that the chosen data accurately describe and fit the research question, thereby improving the final model's effectiveness.

2.1 Data Extraction

This section outlines the process and results of obtaining data from two soccer websites, FBref [9] and Transfermarkt [10], and explains how this data was used to build an XGBoost model. The primary programming tool for extracting this data was a publicly accessible data scraper

[11]. Player-season data from the top five European leagues (2018–2023) were extracted using R, yielding ~17,000 non-distinct player-seasons and ~8,000 observations with Transfermarkt market values. This dataset comprises approximately 300 relevant features, including expected goals, completed tackles, and successful passes, among others. These attributes serve as the foundation for subsequent analyses. The data was stored in CSV files on Google Drive, making it easily accessible via Google Colab.

The dataset for the Injury Duration Model contained 257 features used to predict injury duration. The primary objective of this dataset is to forecast the duration of injuries that players are likely to endure during the subsequent year of their careers, ranging from 0 to 365 days. Notably, the dataset incorporates a ground truth variable, providing a reference point for the accuracy of our predictions.

The dataset for the Player Value Model encompasses the same statistics as before, with the notable addition of a new column representing injury severity, a pivotal component of the model. With an added feature, this dataset will serve as the data frame for the Player Value Model, further enhancing its predictive capacity. Subsequently, the results generated are compared against a ground truth value sourced from Transfermarkt. This step is critical, as it provides an objective means to evaluate the model's accuracy and effectiveness.

2.2. Data Concatenation

After acquiring the statistical and injury histories of players in the top 5 European leagues from 2018 to 2023, the data needed to be concatenated, as it was separated into distinct tables corresponding to different types of statistics, including standard, shooting, passing, defense, playing time, and others. All of the data for the players was concatenated using performance data

from FBref and Transfermarkt. Using the pandas library to read the CSV files and create dataframes, the code combines all the different features based on the players' names for the seasons from 2018 to 2023.

2.3. Data Cleaning

2.3.1 Injury Duration Model

Data cleaning can start after combining all players and their statistics into a large data frame. The first step was to create a model for predicting the duration of injuries for each player. The model was fine-tuned by adjusting the number of preceding years included, with multi-year historical windows serving as an intermediate step in the model development process. A sliding window algorithm was implemented with a parameter, years, that concatenates the stats of the number of years given. For example, if the years parameter were 3, a player, X, would have data from their 2018, 2019, and 2020 seasons concatenated into one row. This process was repeated for each subsequent season (e.g., a player's 2019 row would incorporate statistics from 2020 and 2021). After merging player statistics, a sliding-window design aggregated the preceding y seasons, ranging from 1 to 5, into a single feature vector per target season. For example, with y = 3, the 2018–2020 features predict 2021 outcomes. The window continues to roll forward until the end of the data, 2023. Identifiers such as jersey number and player name were removed as they added no additional information.

2.3.2 Player Value Model

Similar to cleaning the Injury Duration Model, all players and their statistics were combined into a large data frame, after which a sliding window algorithm was implemented. Then, unnecessary columns, including identifiers such as jersey number and player name, were removed, among

others. Five different data frames were created, each corresponding to 1-5 years of the sliding window method.

3. MODEL SELECTION

Choosing the correct model is another crucial step in the process, as each model offers unique advantages tailored to the specific data or task at hand.

3.1 XGBoost model

XGBoost, or XGB, is a decision tree and ensemble algorithm optimized using bagging and boosting [12]. An ensemble algorithm is defined as combining several models to improve overall performance by combining smaller models trained on different subsets of the data, and utilizing bagging and boosting to accomplish these objectives (Hachcham, 2023). In each decision tree, features are split until reaching leaves, the terminal nodes that contain prediction values derived from minimizing loss functions. Bagging decreases the overall variance and bias in a model, minimizes overfitting, and deals with higher-dimensional data more efficiently (Biswal, 2023). Boosting converts weak learners into strong learners, thereby increasing the model's overall predictive accuracy (AWS 2023). Weak learners are equivalent to taking random guesses and are prone to overfitting, which means they cannot accurately classify data that may differ from the original dataset. On the other hand, strong learners exhibit significantly higher predictive accuracy. The data obtained has many features, several of which may contribute little to predicting injuries or player value, and are therefore classified as weak learners. Due to its gradient boosting capabilities, XGB generally outperforms other models in terms of both accuracy and speed (Ruiz, 2023). Since the dataset used had some missing values, sparse data,

especially for injured players, and a large size, XGB was the best fit for analyzing this data, so it was used in this research.

3.2 Model Building

The flowchart below is for reference when discussing the following two models.

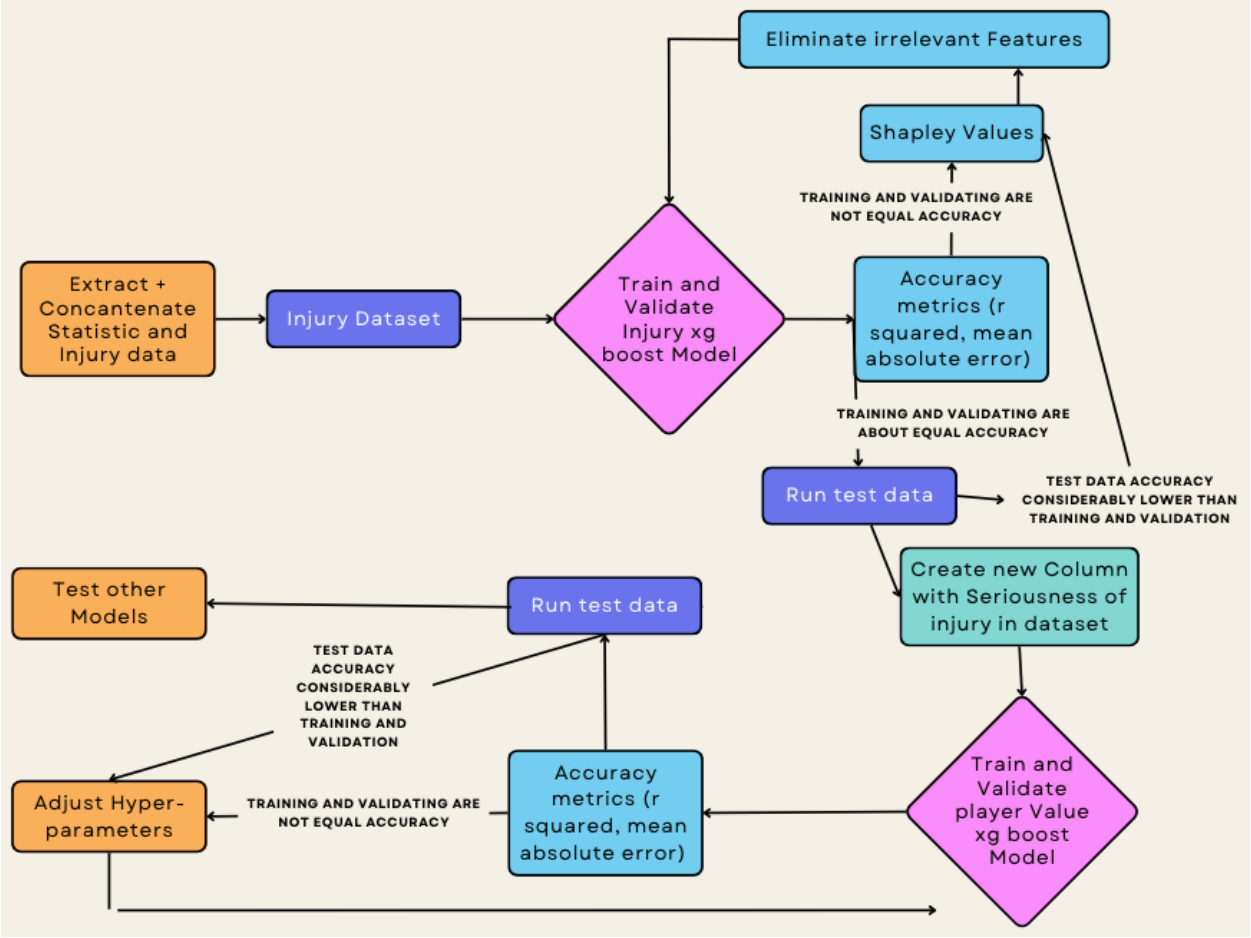

Figure 1: Flowchart for Injury Duration Model and Player Value Model

3.2.1 Injury Duration Model

Five different approaches were used, using one to five years of player statistics to predict injury duration. After cleaning this data, one-hot encoding was applied, a technique used to convert

categorical data into boolean values [13]. The main upside to this approach is the ability to run a default XGBoost model with no hyperparameters tuned using the XGBoost package and scikit-learn, a wrapper for XGBoost, and evaluate the importance of the features. This analysis identifies the features that contribute most to model predictions (Awan, 2023). This allows for further data cleaning by eliminating features that contribute little to the model or debugging features that should not be present in the dataframe, while retaining the features that contribute the most to the model.

Hyperparameter	Description
seed	Allows for less randomness in training a model
max_depth	The maximum depth of the tree: More complex with higher values, but more likely to overfit.
learning_rate	Prevents overfitting by shrinking feature weights
min_child_weight	Determines how conservative the algorithm will be, only producing a child if the sum of instance weight is less than the min_child_weight
gamma	Minimum loss reduction is required to make a further partition on a leaf node of the tree; a larger gamma means a lower chance of overfitting

Table 1: Hyperparameters

3.2.2 Player Value Model

Similar to the Injury Duration Model, five different approaches were used using one to five years of player statistics to predict a player's value. After cleaning and one-hot encoding the data, an XGB model with no hyperparameters was run to evaluate the importance of the features, allowing for further data cleaning. All of the models were trained using an 80% training and 20% testing split. This means that 80% of the data is used to train the model, and the other 20% is used to test the model's accuracy. After eliminating features, the different hyperparameters can be iterated to create the most accurate model. The same hyperparameters were used in both the Injury Duration Model and the Player Value Model, as shown in Table 1.

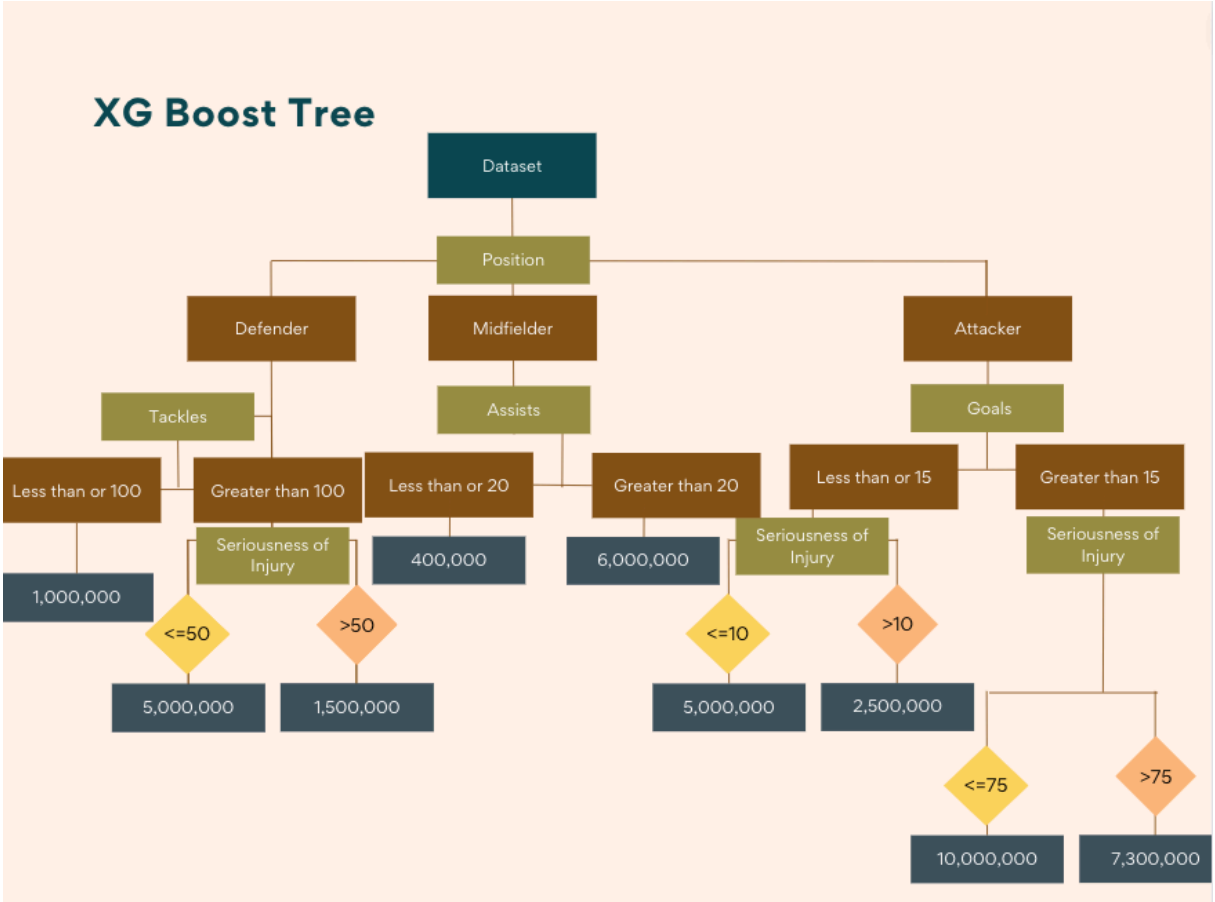

Figure 2: XGBoost Modeling Example

XGBoost will be structured into a decision tree to visualize the Player Value Model. As the tree's depth increases, the relative significance of factors contributing to a player's value diminishes. Notably, pivotal factors such as the player's position, the most pertinent position-specific statistics, and the impact of injuries are likely to occupy prominent positions near the upper tiers of this tree-like visualization.

It is also essential to underscore a fundamental principle in machine learning dataset construction: the balance between the number of rows and columns. In line with best practices, the dataset adheres to the guideline that there should ideally be a minimum of ten times as many rows as columns. This minimum ratio not only ensures the dataset's robustness but also enhances the performance and reliability of subsequent machine-learning models [14].

4. RESULTS

The Injury Duration Model yielded no significant results. No feature was particularly important in the prediction, and the predictions were highly inaccurate, resulting in a negative R-squared (R^2) score. The feature importance scores were all uniform and close to zero, indicating that none of the features contributed significantly to the model's performance. This indicates the model fails to capture the trend in the data, performing worse than a mean-based baseline. As shown in Table 2 below, the negative R^2 score reflects that even after the hyperparameters were optimized, the model was unable to outperform random guesses. Consequently, the Injury Duration Model could not provide a reliable injury-severity estimate, so it could not be integrated into the Player Value Model.

	R^2	Mean Absolute Error (Days)
1 Year	-0.017	46.11
2 Years	-0.166	44.67
3 Years	-0.013	41.71
4 Years	-0.021	49.36
5 Years	-0.403	69.34

Table 2: Results from the Five Injury Duration Models using Sliding Window Algorithm

With the Player Value Model, when the XGBoost model was trained, the default hyperparameters used were tree_method = “approx”, seed = 1, enable_categorical = True, n_estimators = 200, and eval_metric = “mae”. The enable-categorical hyperparameter was set to true since categorical data was used within the dataset, such as player position and country of origin. This lets the model know that categorical data was used and processes it correctly. The hyperparameter n_estimators indicates the number of trees the model should create. If the number is too high, it can cause overfitting, or if too low, it could cause underfitting. At around 200 estimators, the model started to overfit, so 200 was used as the stopping point. Using the default hyperparameters for the year 1 to year 5 models, the model’s accuracy is evaluated using R^2 and Mean Absolute Error (MAE), as shown in Table 3.

	R^2	Mean Absolute Error (€)	Mean Absolute Error ($)
1 Year	0.755	4.60M	5.03M
2 Years	0.816	3.58M	3.91M
3 Years	0.823	4.29M	4.68M
4 Years	0.856	3.50M	3.82M
5 Years	0.816	4.08M	4.46M

Table 3: Results from the Five Player Value Models using Sliding Window Algorithm

To further interpret the Player Value Model, feature importance scores from the XGBoost algorithm were examined. Across all sliding windows, the most influential predictor was the player's market value from the previous season by a large margin. Injury-related variables, such as days missed or severity, consistently did not appear within the top few features, always having a feature importance close to zero. The feature importance analysis highlights which features play the most significant role in predicting a player's value. The feature importance analysis of the best-performing model, the four-year model, is displayed in Table 4.

Feature	Feature Value
Player Market Value from 3rd year	0.321
Non-Penalty Goals in 2nd year	0.071
Plus/Minus per 90 Minutes in 2nd year[1]	0.049
Touches in the Defensive third in 1st year	0.034
Progressive Passes Received in 4th year[2]	0.025

Table 4: Top 5 Features by Importance in Four-Year Player Value Model

5. DISCUSSION

5.1 Findings

This study developed two models, one to predict injury duration and another to forecast player value. The Injury Duration Model yielded no significant results, producing extremely inaccurate predictions, which suggests that the available injury data was too limited to capture meaningful patterns. On the other hand, the Player Value Model achieved a strong performance across all of the sliding windows, with the four-year model being the best due to its high R^2 (0.856) and low MAE ($3.82M). The sliding windows algorithm was implemented to strike a balance between historical data and sample size. The four-year window emerged as the right balance between

[1] Goals scored minus goals allowed by the team while the player was on the pitch per 90 minutes played.

[2] Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch.

historical depth and overly constraining the sample size. This finding can help future researchers design models that consider both temporal depth and sample size. However, the strength of the four-year model reveals the flaws.

When examining the impact of each feature, the primary reason the one-year model underperformed relative to the others was the absence of prior market valuation as a feature. For the other models, the top feature was the player's market value from the previous year. The model effectively reinforces and expands on existing valuations, but exhibits limited capacity to identify undervalued players with minimal or no prior valuation data. In the four-year model, as seen in Table 4, the non-penalty goals, plus/minus per 90 minutes, and progressive passes received are sensible features that contribute to a player's market value. Typically, players who score goals are worth the most compared to other players, as goals are the most important part of soccer. Similarly, plus/minus per 90 minutes measures a player's impact on the goal differential while they are playing. If it is consistently positive, the player is likely to have a significant effect on their team. Although the quality of their squad can confound this metric, it is typically a good measure of the player's individual impact while on the pitch. Finally, the progressive passes received indicate how much the team values the player. If the player consistently receives the ball, they are likely trusted by their teammates and coach to take the ball and make a play, indicating that they are a skilled player.

A key finding of this study was that injury history, which would seem necessary to a player's value, did not enhance the predictive capacity of the models. Rather than disregarding injuries, this study highlights the lack of publicly available injury data, which reduces more complex data, such as recovery processes and recurrences, to a single data point. Without more nuanced data, the model is unable to account for how injuries can truly impact a player's trajectory. A

predictive Injury Duration Model could still be implemented, but the data would need to be much larger and in-depth.

5.2 Limitations

Several limitations must be noted that may have impacted the accuracy of the model built. First, the quality of the injury data was a significant challenge. As mentioned, most clubs are reluctant to publicly disclose data on their players, which in turn limits the publicly available data and reduces the predictive power of the Injury Duration Model. Another limitation is the Player Value Model's reliance on past market value, which risks predictions being powerfully shaped by existing valuations rather than uncovering hidden insights. Additionally, the scope of the dataset, encompassing six seasons in the top five European leagues, may limit the applicability of these findings to youth leagues and academies, which limits the model's ability to identify undervalued or overlooked players.

5.3 Future Research and Implementation

Future research should address these limitations by incorporating larger and more in-depth injury datasets, with extensive medical information about each player and injury. Testing other machine learning or AI models, such as recurrent neural networks, may also be more effective at capturing the sequential nature of player development and injury risk. To further generalize the model, data should be included from academies and other leagues outside the top five European leagues. Overall, this research highlights the potential benefits that football clubs can gain by incorporating advanced machine learning into their everyday decision-making processes. This model can nevertheless serve as a secondary evaluation layer, supporting scouts in identifying potential risks, validating assessments, or flagging players whose valuations diverge from

statistical and injury-based indicators. As data collection improves and more generalized models emerge, these models will complement scouting teams to reduce financial risk and enhance decision-making in the soccer transfer market.

6. CONCLUSION

This study examined whether publicly available injury histories improve machine-learning forecasts of soccer player market value. Across one to five-year windows, the four-year model achieved the best out-of-sample performance ($R^2 = 0.856$; MAE ≈ $3.82 million). However, injury features contributed little additional signal relative to lagged market value and core performance statistics, likely reflecting the coarse granularity of public injury data and potential embedding of injury effects within existing market valuations. These findings suggest that meaningful integration of injuries into valuation will require richer medical and recovery information, time-aware modeling, and position-specific approaches. As broader datasets become available, such models can complement scouting by quantifying risk and value with greater transparency.

Bibliography

1. Tierney, Martin. *Too big to go down? A study comparing the price inflation of Premier League football players with the characteristics of economic bubbles*. Diss. Dublin, National College of Ireland, 2020.

2. Li, Chenyao, Stylianos Kampakis, and Philip Treleaven. "Machine learning modeling to evaluate the value of football players." arXiv preprint arXiv:2207.11361 (2022).

3. Zhang, D., & Kang, C. (2021). Players' value prediction based on machine learning method. *Journal of Physics: Conference Series*, *1865*(4), 042016. https://doi.org/10.1088/1742-6596/1865/4/042016

4. Franceschi, M., Brocard, J., Follert, F., & Gouguet, J. (2023). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12552

5. Majumdar, A., Bakirov, R., Hodges, D., Scott, S., & Rees, T. (2022). Machine learning for understanding and predicting injuries in football. *Sports Medicine - Open*, *8*(1). https://doi.org/10.1186/s40798-022-00465-4

6. Hamilton, G. M., Meeuwisse, W. H., Emery, C. A., Steele, R. J., & Shrier, I. (2011). Past injury as a risk factor: An illustrative example where appearances are deceiving. *American Journal of Epidemiology*, *173*(8), 941–948. https://doi.org/10.1093/aje/kwq461

7. Ekstrand, J., Hagglund, M., & Walden, M. (2009). Injury incidence and injury patterns in professional football: The UEFA Injury Study. *British Journal of Sports Medicine*, *45*(7), 553–558. https://doi.org/10.1136/bjsm.2009.060582

8. Eliakim, E., Morgulev, E., Lidor, R., & Meckel, Y. (2020). Estimation of injury costs: Financial damage of English Premier League teams' underachievement due to injuries. *BMJ Open Sport & Exercise Medicine*, *6*(1). https://doi.org/10.1136/bmjsem-2019-000675

9. https://fbref.com/

10. https://www.transfermarkt.us/

11. https://github.com/JaseZiv/worldfootballR_data

12. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

13. Samuels, Jamell. (2024). One-Hot Encoding and Two-Hot Encoding: An Introduction. 10.13140/RG.2.2.21459.76327.

14. *How much data is needed for machine learning?*. Graphite Note. (2023, September 8). https://graphite-note.com/how-much-data-is-needed-for-machine-learning