**Important website:** [**https://fbref.com/en/comps/Big5/misc/players/Big-5-European-Leagues-Stats**](https://fbref.com/en/comps/Big5/misc/players/Big-5-European-Leagues-Stats)

[**https://fbref.com/en/expected-goals-model-explained/**](https://fbref.com/en/expected-goals-model-explained/)

**A Novel Model for Forecasting Soccer Player Value Incorporating Risk Factors through Machine Learning**

**Question: How do different statistics and injuries affect a soccer player’s value in the transfer market.**

Intro/Overview -

Effectively assessing the value of a soccer player, particularly gauging the potential risks associated with their acquisition, has posed a considerable challenge for numerous sports teams. Over the recent months, soccer clubs have made substantial financial investments in player acquisitions, only to find that these signings either didn’t help their team or lost their team money. This issue has become increasingly prevalent. While several machine learning models incorporate statistical metrics such as goals scored, assists provided, and red cards received to predict a player's value, they often overlook a critical factor: injuries [1,5,6]. While articles have focused on just injuries, only a few have incorporated injuries in player valuation [7,8].

Injuries play a pivotal role in shaping a soccer player's career trajectory and influencing their future performance. A comprehensive study spanning the years 2001 to 2008, encompassing 23 professional soccer teams across Europe, revealed that each player sustained approximately 2.0 injuries per season on average. Notably, re-injuries accounted for approximately 12% of these injuries, resulting in longer periods of absence compared to initial injuries [2]. Furthermore, injuries exact a substantial economic toll on English Premier League teams, with an estimated annual loss of approximately 45 million pounds sterling [3].

My primary objective is to develop a precise and dependable system for predicting the value of a soccer player. This predictive model aims to assist professional soccer coaches and scouts in assessing whether a player is a suitable fit for their team. It will revolutionize player evaluation by factoring in not only their historical statistics and performance metrics but also their injury history. This multifaceted approach will enable clubs to identify undervalued players with significant potential and determine optimal times to sell existing assets.

Additionally, a robust predictive model can highlight various risk factors associated with a player, such as their injury susceptibility, age, and prior performance history. By leveraging such insights, clubs can proactively manage their player assets, make informed decisions regarding player acquisitions, and potentially identify performance trends that may lead to increased

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 within the realm of professional soccer.

Methods -

Python libraries:

Xgboost (machine learning algorithm), pandas, pyreadr (reading files), numpy, sklearn/scikit

Google.colab library (integrates google drive with colab, storage of files, dataframes, models)

R packages: JaseZiv/worldfootballR

Using a combination of Google Colab and IntelliJ Jupyter Notebooks to build models and RStudio to extract data

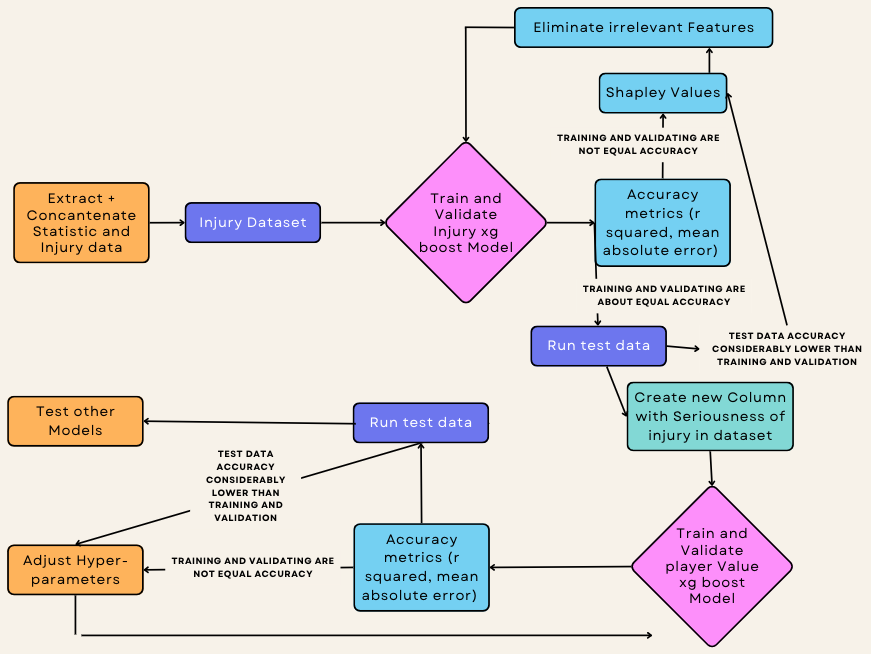
Obtaining a dataset:

https://github.com/JaseZiv/worldfootballR\_data I have collected extensive data from sports statistics websites, specifically utilizing the publicly accessible APIs and packages provided the GitHub above, giving me access to FBref and transfermarkt. This data extraction process enabled me to compile a comprehensive dataset spanning the seasons from 2018 to 2023, encompassing professional European players who were active during this time frame.

Using R, I extracted all available data for these seasons, resulting in a dataset of approximately 13,000 players. Among these players, approximately 8,000 have market values assigned, rendering them particularly significant for my analytical purposes.

This dataset includes around 300 pertinent features such as expected goals, tackles completed, and successful passes, among others. These attributes serve as the foundation for my subsequent analyses.

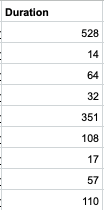
Notably, it's essential to underscore a key principle in machine learning dataset construction: the balance between the number of rows and columns. In line with best practices, my dataset adheres to the guideline that there should ideally be a minimum of ten times the number of rows compared to columns. This optimal ratio not only ensures the robustness of the dataset but also enhances the performance and reliability of subsequent machine learning models [4].



This flowchart is for reference when talking about the following two models.

Dataset for Injury Model:





Presented here is a condensed glimpse of my statistical dataset, meticulously curated for the purpose of predicting injury duration – a critical indicator of injury severity. This dataset comprises 214 distinctive features and spans N individual players, forming the basis of my analytical endeavor.

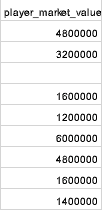
The primary objective of this dataset is to forecast the duration of injuries that players are likely to endure during the subsequent year of their careers, a timeframe ranging from 0 to 365 days. Notably, the dataset incorporates a ground truth variable, providing a reference point for the accuracy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

Following the model's analysis and prediction outcomes, a comprehensive examination of Shapley values will be undertaken. This examination will unveil the features that exert the most significant influence (most correlated with) on the prediction of injury duration. Subsequently, adjustments to the model will be made, optimizing its performance and predictive accuracy.

Additionally, I will have the flexibility to fine-tune the number of preceding years (e.g. selecting a range such as three years of historical injury data) that inform the prediction of injuries in the upcoming year of a player's career. This adaptability is vital, as the choice of this parameter can significantly impact the model's performance. Setting it too high may yield a more accurate model but could result in limited available data, potentially compromising the model's robustness. Thus, striking the right balance is crucial to achieving both precision and data sufficiency.

Dataset for Player Value Model:



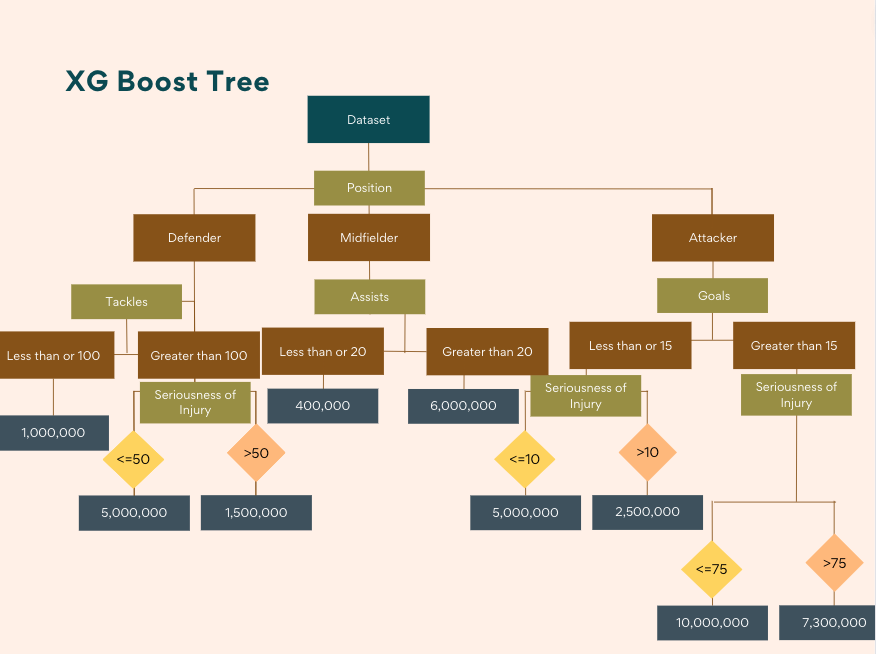


The dataset in question encompasses identical statistics as before, with the notable addition of a new column representing the injury severity, a pivotal component of our injury model. This dataset, now enriched with M+1 features, will serve as input for the Player Value model, further enhancing its predictive capacity. Within this dataset, N individual players contribute their insights and data.

Subsequently, the results generated by the Player Value model be compared against a ground truth value sourced from the transfermarket website. This step serves as a critical measure of the model's accuracy and effectiveness.

Upon reviewing the model's performance, adjustments to hyperparameters will be contemplated. Parameters such as maximum depth, learning rate, maximum leaves, among others, will be fine-tuned to optimize the model's predictive capabilities and ensure its alignment with the actual player values. This iterative refinement process represents a vital aspect of enhancing the model's precision and overall performance.

The dataset will have the same statistics with one additional column that the injury model will give (seriousness of injury). This is a shortened dataset that I will feed into the Player Value model. I will have M+1 features because of the added feature of injury duration and N players. The output of the player value model will be compared to a truth value from the transfermarket website. After looking at the accuracy of this model, I can adjust the hyperparameters such as max depth, learning rate, max leaves, etc.



To visualize my Player Value model, xgboost will be structed into a tree. 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.

* **what is your train-validation-testing split (percentage of data in each and how you split them, i.e., was it a random split)?**
* **what hyperparameters / metrics are you using to tune the XGboost model?**
* repeat these items for the player value model.

Presentation of Data -

Alternate Perspectives/expansion and restate thesis -

Conclusion -

Bibliography

1. Li, C., Kampakis, S., & Treleaven, P. (n.d.). Machine learning modeling to evaluate the value of football players - arxiv.org. https://arxiv.org/pdf/2207.11361.pdf
2. 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
3. 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 &amp; Exercise Medicine*, *6*(1). https://doi.org/10.1136/bmjsem-2019-000675
4. *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
5. 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
6. 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
7. 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
8. 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