Research Article

Predicting Injury Recurrence and Recovery Patterns

Victor Pastore Marqueti [1]*

[1] student; https://github.com/victorpmarqueti/senior_project_soccer.git

Abstract

Injury recurrence is a persistent issue in professional sports, affecting athlete performance, long-term health, and overall team success. This project explored patterns of injury recurrence among professional soccer players and examined how machine learning models can help in predicting injury outcomes. Using a dataset of player injury histories, the project approached two predictive tasks: estimating the number of days until the recurrence of the same injury (regression) and predicting whether an injury would recur (classification). Several models were tested, with Ridge Regression and Gradient Boosting emerging as top performers. Gradient Boosting showed the strongest performance across both tasks, with high predictive accuracy and recall, while Ridge Regression was particularly effective in estimating recurrence timing. These findings indicate that machine learning can offer valuable insights into injury patterns and recurrence risks. In future work, access to larger and more comprehensive datasets could further enhance model performance and support improved player health and team management strategies in professional soccer.

Keywords: injury recurrence; machine learning; regression; classification; soccer analytics

1. Introduction

Soccer is one of the most globally followed and economically impactful sports, with millions of athletes participating at both professional and amateur levels. Injuries are an unavoidable aspect of professional soccer, significantly impacting athletes careers, long-term health, and team performance. Recurrent injuries are particularly concerning, as they can lead to prolonged absences, reduced performance levels, and increased healthcare costs. Studies have shown that approximately 12% to 30% of all sports injuries are recurrences of previous ones, highlighting the need for better prevention and prediction strategies [1,2,7].

Understanding how and when injuries recur is essential for improving player well-being, optimizing training and recovery protocols, and informing medical and coaching staff decisions. With the increasing availability of structured sports data, data science methods have become a vital tool in sports analytics, offering new ways to identify injury risks and manage athlete health more proactively [3].

While athlete injuries have been extensively studied and it is being discuss across the globe in sports medicine, analyzing recurrence patterns using player-specific historical data is an area with untapped potential[4]. This project aims to identify which injury types are most likely to recur, how long it typically takes for recurrence, and which player

characteristics—such as age, playing position, and recovery duration—are associated with elevated recurrence risks.

To address these questions, we applied various machine learning models to a curated dataset of professional soccer player injuries. Our goal is to generate interpretable predictions and visualizations that can aid in injury management strategies. This project contributes to the growing field of health analytics in sports and opens the door for future work in more personalized and data-driven approaches to athlete care.

2. Data

The data for this project was extract using programming language R, using the worldfootballR package [5]. This package allows access to publicly available football data from websites like Transfermarkt[6]. Using this tool, extracted detailed injury histories for a range of professional soccer players. The dataset includes variables such as player names, age at injury, injury type, position, number of games missed, and injury duration. To allow for the analysis of injury recurrence, each injury was linked to a date, enabling the calculation of how much time passed between injuries of the same type for the same player.

To prepare the dataset, cleaning steps were made. Injury descriptions were format into categories to reduce inconsistencies (e.g., grouping similar ankle related injuries). Dates were converted to correct formats, and the recovery duration column was cleaned by removing text like "days" and converting the values into numeric format, also removed entries with missing or invalid date information.

Injuries related to COVID-19 or quarantine were removed from the dataset, as we still do not fully understand how such cases might impact injury recovery or recurrence patterns. These situations could introduce unclear or misleading factors into the analysis. Until more is known about their long-term effects, we decided to exclude them to maintain focus on traditional injury patterns.

Two target variables were made:

- days_until_recurrence: a numeric value representing the days until the same injury type recurred for the same player.
- recur_same_injury: a binary value indicating whether the same injury happened again (1) or not (0).

Figure 1 shows the injury categories with the highest recurrence rates. Muscle, hamstring, and knee-related injuries were the most frequently recurring, suggesting they may require extra attention from medical staff.

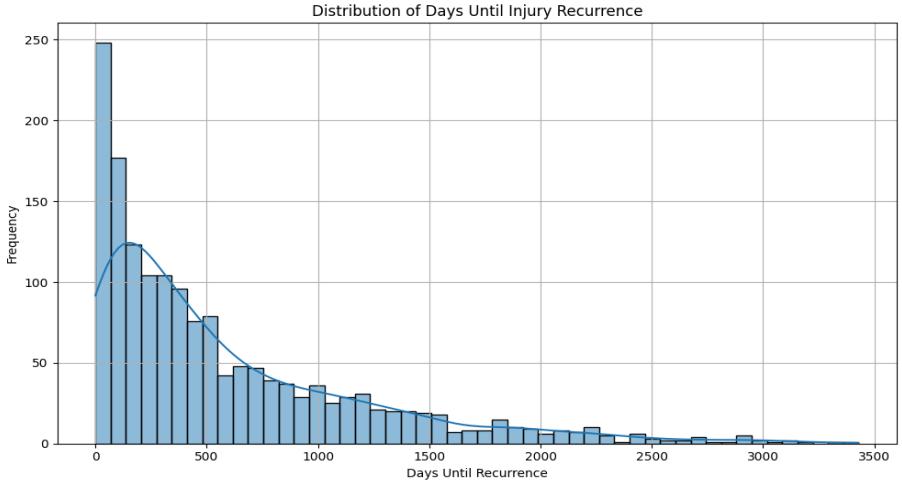

Figure 1. Recurrence rate by injury category.

Figure 2 visualizes the distribution of days until recurrence across different injury categories. Some injuries, like muscle and hamstring injuries, tend to recur sooner than others.

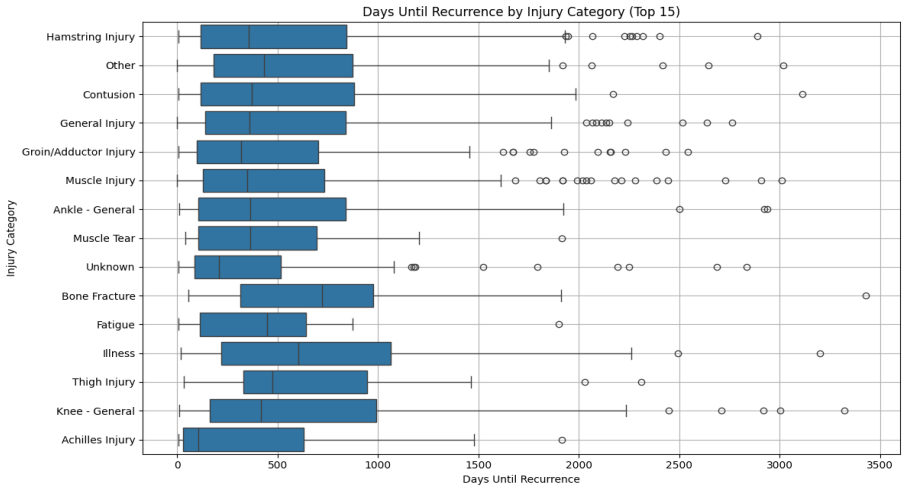

Figure 2. Days until recurrence by injury category.

Figure 3 shows the overall distribution of recurrence intervals. The data is highly skewed, with most recurrences happening within the first 500 days.

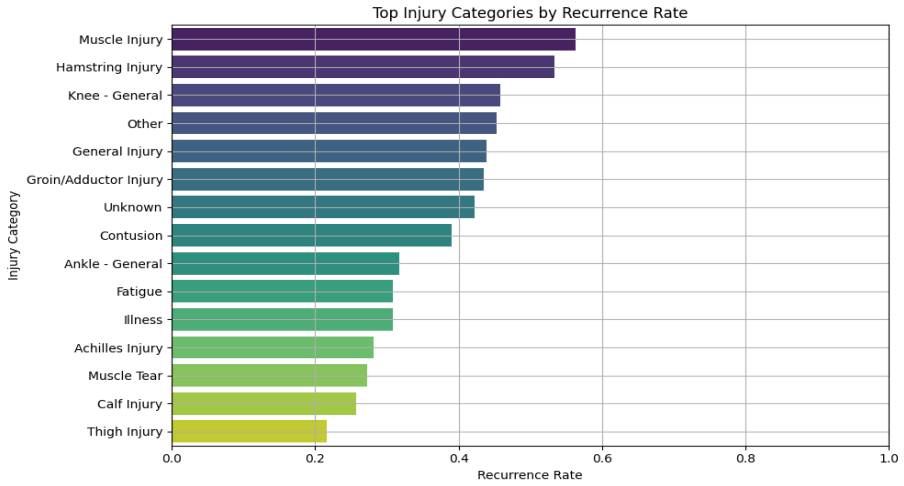

Figure 3. Distribution of recurrence intervals.

3. Methods

To analyze the recurrence of injuries, two different machine learning problem types were defined: a regression to estimate the number of days until recurrence, and a classification to predict whether an injury would recur or not.

For the regression target, several models were tested, however some models had better results and were use in the continuity of the project, including:

- Linear Regression, used as a baseline for its simplicity and interpretability.
- Ridge Regression and Lasso Regression, to improve generalization by adding regularization and managing correlated features.

- Gradient Boosting Regressor, selected for its ability to capture complex, non-linear relationships.

For the classification target, the following models were used for the project

- Logistic Regression, serving as a baseline model.
- Gradient Boosting Classifier, for its strength in predictive accuracy.
- AdaBoost Classifier, which improves by focusing on previously misclassified cases.
- LightGBM Classifier, chosen for its efficiency and strong performance on larger datasets.

Evaluation metrics were chosen according to the task type. For regression models, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were used. For classification, Accuracy, Precision, Recall, and F1-Score helped assess the quality of predictions, especially in relation to class balance and the cost of misclassifications.

This selection process helped focus on models that not only performed better but also provided insight into the recurrence of athlete injuries in professional soccer.

4. Results

This section presents the performance results of the classification and regression models applied to the injury dataset. Evaluating the metrics and make visualizations to each model's predictive ability and interpret their effectiveness in the context.

4.1 Classification Results

To predict whether an injury would recur, tested four classification models: Logistic Regression, Gradient Boosting, AdaBoost, and LightGBM. These models were evaluated using classification metrics...

Table 1 displays the evaluation scores for each classification model. Gradient Boosting achieved the strongest performance across all metrics, including the highest accuracy (0.82), precision (0.84), recall (0.94), and F1 score (0.88). Logistic Regression and LightGBM also showed solid performance, especially in recall and F1 score. AdaBoost, while effective at identifying positives (recall = 0.83), had lower accuracy and AUC, indicating less balanced predictions overall.

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic	0.79	0.82	0.89	0.85	0.68
Gradient	0.82	0.84	0.94	0.88	0.69
AdaBoost	0.74	0.81	0.83	0.82	0.62
LightGBM	0.76	0.83	0.89	0.86	0.67

Table 1. Summarizes the evaluation metrics for all models.

Figure 4 presents the ROC curves for the four classifiers. Gradient Boosting achieved the best AUC (0.69), followed closely by Logistic Regression (0.68) and LightGBM (0.67), while AdaBoost lagged (0.62). The ROC curve illustrates that Gradient Boosting consistently provided the best trade-off between sensitivity and specificity.

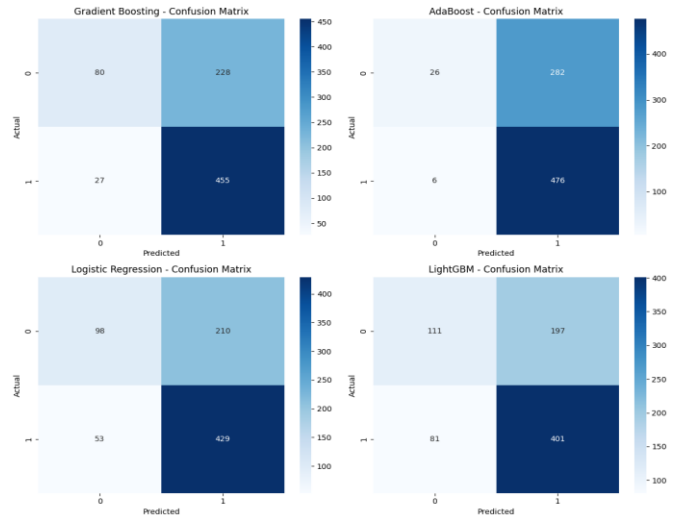

Figure 4. Confusion Matrix for All Models.

Figure 5 shows the confusion matrices for all models, revealing how they classified true positives and false positives. Gradient Boosting had the lowest number of false negatives (27), indicating strong ability in identifying recurring injuries. In contrast, AdaBoost and LightGBM misclassified a greater number of cases, contributing to their lower overall

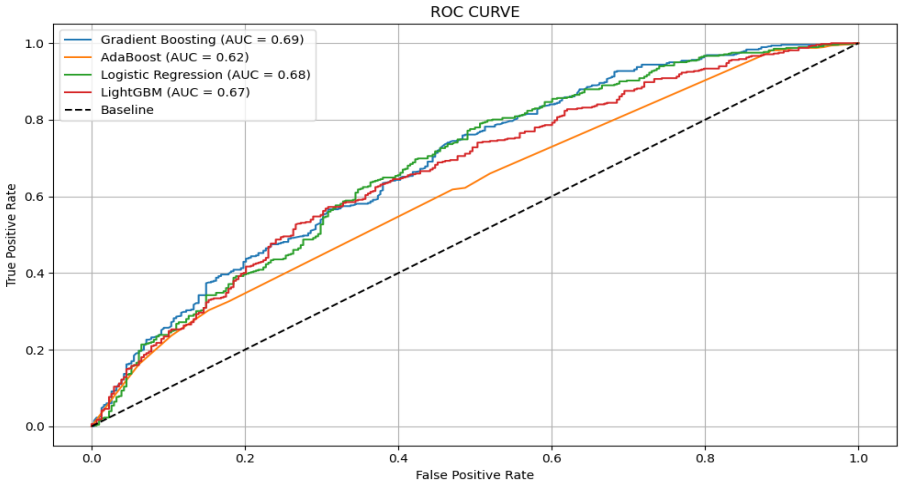

Figure 5. Roc Curve for All Models.

Gradient Boosting delivered the strongest classification performance, achieving the highest scores across key evaluation metrics and demonstrating consistent reliability in predicting injury recurrence.

4.2 Regression Results

To estimate the number of days until the recurrence of the same injury, four regression models were tested: Linear Regression, Ridge Regression, Lasso Regression, and Gradient Boosting Regressor. The models were evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE).

Lasso Regression and Ridge Regression delivered the strongest performance, with Lasso achieving the lowest RMSE (600.19) and MAE (468.97). Ridge Regression followed closely, nearly matching Lasso's performance. These results suggest that regularized linear models are well suited for this regression task. Linear Regression also performed

reasonably, while Gradient Boosting showed slightly higher error values (RMSE of 624.34 and MAE of 480.26), indicating less consistency in its predictions.

Figure 6 displays predicted versus actual recurrence days. All models tended to underpredict longer recurrence intervals, although Lasso and Ridge aligned more closely with the ideal diagonal line.

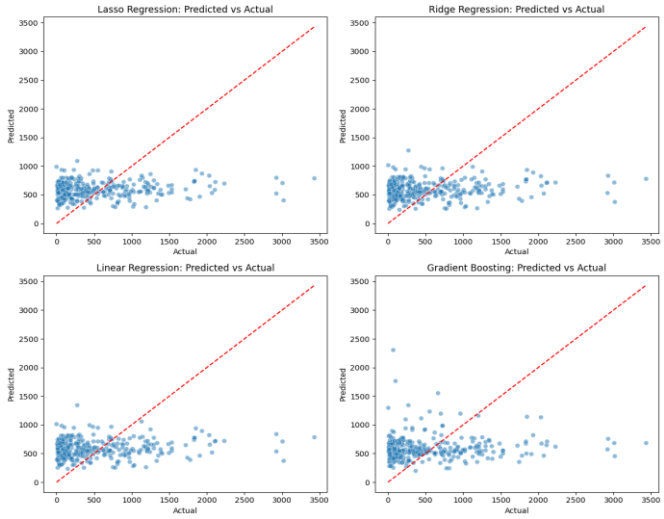

Figure 6. Predicted vs Actual Scatter Plots for All Models.

Figure 7 presents a boxplot of the prediction errors, where Lasso and Ridge showed narrower interquartile ranges, suggesting more consistent predictions.

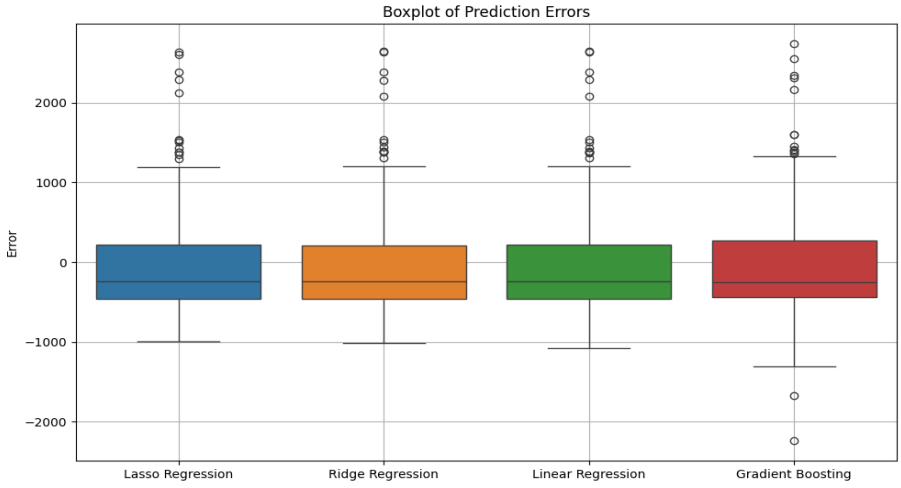

Figure 7. Boxplot of Prediction Errors.

Figure 8 shows residual plots to visualize model bias. Ridge and Lasso maintained relatively stable residuals around zero, whereas Gradient Boosting displayed a wider spread, especially for higher recurrence values.

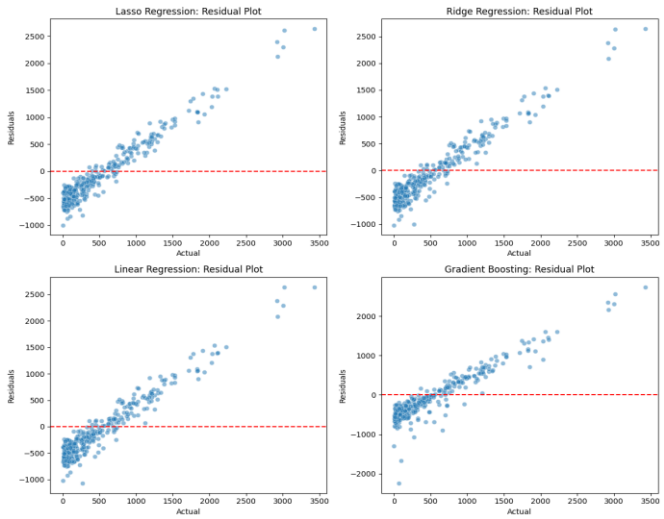

Figure 8. Residual Plots for All Models.

Overall, Lasso and Ridge Regression delivered the most reliable results, suggesting that linear models with regularization are effective for estimating recurrence timing. Although Gradient Boosting introduced more variability, it may still offer advantages with further tuning or additional features.

5. Discussion

The classification results I believed that showed that Gradient Boosting and Ridge Regression achieved the best overall performance in predicting injury recurrence outcomes. Gradient Boosting excelled in the classification task, delivering high recall and strong accuracy in identifying whether an injury would recur. In the regression task, Ridge Regression performed best in estimating the number of days until recurrence, with Lasso Regression showing similarly strong performance.

The regression models faced greater challenges when estimating the exact number of days until recurrence. Ridge Regression performed better than other models, the overall prediction errors suggest that modeling recurrence timing remains complex. I believed that it is due to the absence of variables that influence recovery and reinjury risk.

These findings illustrate both the promise and the current limitations of applying machine learning to injury analytics in professional sports, not just soccer. While the models show potential in supporting injury management decisions, the results show the need for better and more robust datasets to achieve a more accurate predictions.

6. Conclusions

The project demonstrated the potential of machine learning in analyzing player injury recurrence using historical data. With findings support the growing role of data-driven approaches in athlete health management, offering tools that could assist medical and coaching staff in monitoring injury risks more effectively. However, limitations related to dataset size and the absence of key contextual variables such as training load, rehabilitation protocols, and match conditions, highlight the need for richer, more comprehensive data, if we want to achieve a better result.

Future work should prioritize expanding the dataset to include a wider range of player-specific features and injuries across multiple teams and seasons. Doing so will help improve model robustness, prediction accuracy, and ultimately, the practical impact of machine learning in professional sports injury prevention and recovery planning.

References

1. Ekstrand J, Hägglund M, Waldén M, Injury incidence and injury patterns in professional football: the UEF. injury study, British Journal of Sports Medicine 2011;45:553-558.
2. Junge, A., & Dvorak, J. "Injury surveillance in the World Football Tournaments 1998–2012." British Journal of Sports Medicine 47.12 (2013): 782–788.
3. Chomistek, Andrea K.; Bassett, David R. Jr.. Response. Medicine & Science in Sports & Exercise 50(4):p 877, April 2018. | DOI: 10.1249/MSS.0000000000001510
4. Majumdar, A., Bakirov, R., Hodges, D. et al. Machine Learning for Understanding and Predicting Injuries in Football. Sports Med - Open 8, 73 (2022). https://doi.org/10.1186/s40798-022-00465-4
5. Wragg, J. (2023). worldfootballR: Data from the World of Football (Soccer). Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/worldfootballR/readme/README.html
6. Transfermarkt. Football statistics & transfer news. https://www.transfermarkt.com/
7. Sportsepreneur. (2021, May 3). *Why Is Soccer the Most Popular Sport in the World?* https://sportsepreneur.com/worlds-most-popular-sport/
8. Van Eetvelde, Hans et al. "Machine learning methods in sport injury prediction and prevention: a systematic review." Journal of experimental orthopaedics vol. 8,1 27. 14 Apr. 2021, doi:10.1186/s40634-021-00346-x
9. Windt J, Gabbett TJ, How do training and competition workloads relate to injury? The workload—injury aetiology model, British Journal of Sports Medicine 2017;51:428-435.