Entrepreneurship and Innovation
2025, VOL. 9, NO. 2, 31-36
DOI: 10.47297/wspeiWSP2516-253506.20250902

Research on athlete injury prediction model based on Logistic

Fangyong Wang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, Yunnan, 650221, China

ABSTRACT
In today's sports field, the prevention and management of athletes' injuries have become crucial. Timely and accurate prediction of athletes' injury risk can not only help the team to make a reasonable training plan, but also greatly reduce the loss caused by injury. In this study, based on injury_data datasets, a model of athlete injury prediction was constructed by using Logistic regression model, aiming to achieve the binary classification prediction of athlete injury risk. The injury_data datasets contains several key pieces of information about athletes, such as type of exercise, intensity of training, prior injury history, etc., which are critical to predicting whether an athlete is at risk of injury. We ensure that the data used for modeling is of high quality and accuracy through elaborate pre-processing steps, including missing value processing and outlier detection. In the process of constructing the Logistic regression model, Maximum Likelihood Estimation (MLE) is used to determine the optimal model parameters. In order to evaluate the performance of the model, we use the likelihood ratio test as the evaluation criterion of model fit degree. The experimental results show that our model can effectively predict the injury risk of athletes, and its prediction accuracy meets the preliminary application requirements. Through this study, we have confirmed the feasibility and effectiveness of the application of Logistic regression model in the field of athlete injury prediction, and hope that this method can provide technical reference for subsequent research in similar fields. Future work will focus on introducing more data features and exploring more advanced prediction models to further improve the accuracy of predictions and help optimize athlete health management and injury prevention strategies.

KEYWORDS
Logistic regression; Maximum likelihood estimation; Hypothesis testing; injury_datasets

1 Introduction

In competitive sports, the health and safety of athletes are paramount. Injuries not only impact an athlete's career but can also affect the overall performance of the team. As Leckey et al. (2024) pointed out in their scoping review, the application of machine learning approaches in predicting sports injuries has shown great potential in enhancing injury prevention strategies[1]. As the intensity of training increases, the risk of injuries among athletes also rises, making injury prediction and prevention a key task for sports organizations and coaching teams. Traditional injury prediction methods rely on the experience of doctors and cannot comprehensively cover potential risks, especially hidden or early-stage injuries. However, recent studies have shown that integrating physical fitness tests and body composition data can improve the accuracy of injury prediction. For example, Martins et al. (2022) developed a predictive model for elite football players based on these factors[2]. Modern technologies, particularly data analysis and artificial intelligence (AI), provide new solutions for predicting athlete injuries. By combining algorithms such as Logistic Regression, accurate injury risk assessments can be made.

The prediction model based on Logistic Regression not only improves the accuracy of injury prediction but also provides coaches with scientific data to support the formulation of reasonable training plans, thereby reducing athletes' injury risks. The application of the model also helps sports medicine teams to identify high-risk individuals in advance and develop personalized protective strategies[5]. Overall, as the level of competitive sports increases, the combination of modern computing technology and sports medicine will push injury prediction to higher levels, bringing wide-ranging social and economic benefits to the sports industry[6].

This study aims to analyze athlete injury data through a Logistic Regression model, developing an efficient and accurate injury risk prediction system to support athlete health management and injury prevention. The structure of the paper includes an introduction, literature review, dataset description, methodology section, experimental analysis, discussion, and conclusion, providing a detailed explanation of the research background, methods, experimental results, and future directions for improvement.

2 Related Work

The field of athlete injury prediction has widely applied statistical and machine learning models, including linear regression, artificial neural networks (ANN), support vector machines (SVM), and ensemble learning methods. The linear regression model, due to its simplicity and intuitiveness, is often used for preliminary analysis of injury data. However, injury occurrences often involve multiple factors and nonlinear relationships, and linear regression has limitations when dealing with these complex relationships. Artificial neural networks (ANN) perform exceptionally well in capturing complex data patterns and are particularly suited for handling large-scale athlete physiological and activity data. However, their "black-box" nature limits the transparency and interpretability of the model, especially in scenarios where clear results are required. Support vector machines (SVM) are powerful classification tools that improve classification accuracy by maximizing the distance between data points and the decision boundary. Although SVM can handle high-dimensional datasets, selecting the appropriate kernel function and tuning parameters can be complex for beginners. Ensemble learning methods, such as random forests (RF) and gradient boosting machines (GBM), enhance prediction performance by combining multiple models. These methods exhibit excellent generalization capabilities and robustness to noise but increase computational demands and model complexity.

Among these methods, the Logistic regression model performs well when dealing with binary classification problems (such as injury occurrence). However, some studies have shown that machine learning methods can outperform traditional logistic regression in certain contexts. For instance, Luu et al. (2020) demonstrated that machine learning models were more effective than logistic regression in predicting NHL player injuries[3]. It has the advantages of high computational efficiency and ease of result interpretation, making it useful for providing injury risk predictions for athletes and coaches. While Logistic regression has less flexibility in handling nonlinear relationships, the probability estimates it provides are crucial for decision-making and can help assess athletes' injury risks.

However, Logistic regression also has limitations, especially when the feature space is too high-dimensional or the sample size is small, which may lead to overfitting. To address this issue, researchers often use techniques such as feature selection, regularization, and model diagnostics to improve the reliability of the model.

This study chose the Logistic regression model, specifically using a model with $\log\left(\frac{p}{1-p}\right)=\eta=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$ features, to develop an athlete injury risk prediction system. By using this model, data-driven injury prediction and prevention strategies can be provided for athletes and coaches.

3 Methods

3.1 Logistic Regression Model

Logistic regression is a statistical model used for binary classification tasks. Its core idea is to map the output value of linear regression to the range (0,1) through the nonlinear sigmoid (logistic) function, thereby representing the probability of a specific category. In this study, a logistic regression model is employed to predict athletes' injury risks, aiming to compute the probability of an athlete being injured (1) or not injured (0).

Logistic regression extends linear regression by introducing the sigmoid function, which transforms the linear combination of input features into a probability value. Given a feature vector $X=(x_1,x_2,\ldots,x_n)^T$, where X in R^n, the goal is to predict a binary label Y, which can take values of 0 or 1. The model estimates the conditional probability P(Y=1|X), which can be expressed as[7]:

$$P(Y=1|X)=\sigma(\theta^TX)=\frac{1}{1+e^{-(\theta^TX)}}$$

where $\sigma(z)=\frac{1}{1+e^{-z}}$ represents the sigmoid function, mapping the linear combination results to the (0, 1) interval. The parameter vector $\theta=(\theta_0,\theta_1,\ldots,\theta_n)^T$, which determines the threshold when all feature values are zero, and weights θ_i for x_i(i> 0), which indicate the importance and direction of each feature in the prediction.

Additionally, the logit function (log-odds function) for logistic regression is given by:

$$logit(p)=log(\frac{p}{1-p})=\theta^TX$$

where p represents the probability of an injury occurring. The logit function transforms probability values into a linear relationship, making it suitable for parameter optimization using maximum likelihood estimation (MLE).

3.2 Model Training and Optimization

To train the logistic regression model, the optimal parameter θ needs to be determined. This is typically achieved by maximizing the likelihood function (MLE), often done by maximizing its log-likelihood form. The log-likelihood function for logistic regression is given by:

$$L(\theta)=\sum(Y_ilog(p_i)+(1-Y_i)log(1-p_i))$$

where $P_i=\sigma(\theta^TX_i)$ represents the predicted probability of injury given X_i, and Y_i is the observed value. The optimal parameter θ is found by maximizing.

L(θ), usually through optimization algorithms such as gradient descent. The gradient update rule is $\theta_j := \theta_j - \alpha \frac{\partial L(\theta)}{\partial \theta_j}$, where α is the learning rate, controlling the step size of parameter updates. The loss function is defined as the negative log-likelihood function:

J(θ)=-L(θ)

Gradient descent optimizesθby iteratively adjusting parameters to minimize **J(θ)**, thereby training the model. The choice of optimization algorithm impacts the convergence rate and stability of the model[8]. Common algorithms include:

Batch Gradient Descent (BGD): Uses the entire dataset for each update, ensuring stability but requiring high computational power. Stochastic Gradient Descent (SGD): Updates parameters after each sample, reducing computation but increasing noise in updates. Advanced Variants (Adam, RMSprop, etc.): Combine momentum-based adjustments to enhance convergence efficiency.

Parameter selection involves choosing the learning rate α and incorporating regularization terms (L1 or L2 regularization) to prevent overfitting. The learning rate is often fine-tuned through experiments, starting with a larger value and gradually decreasing to avoid oscillations.

Additional model optimization strategies include: Feature Scaling: Standardizing features to accelerate training convergence. Learning Rate Decay: Dynamically adjusting$\alpha$to balance exploration and convergence. Momentum Techniques: Smoothing updates to prevent drastic changes in parameter values.

Training stops when the loss function converges, i.e., when its change falls below a predefined threshold. The final optimized parameters θ are then used for injury prediction on new athlete data. By integrating statistical methods such as MLE for parameter estimation, hypothesis testing for model evaluation, and regularization for overfitting control, the logistic regression model ensures robustness and interpretability in predicting athlete injuries.

4 Experiment

4.1 Dataset Description

The athlete injury prediction dataset (injury_data) comes from the Hejing community[13]. This dataset is a comprehensive athlete injury analysis dataset composed of 1000 simulated data records. It includes information about athletes' basic physical characteristics and training conditions, such as player age, weight, height, previous injuries, training intensity, and the correlation of each parameter in predicting the likelihood of injury. The following table outlines the fields contained in the dataset and their corresponding explanations, providing a foundation for understanding the dataset structure and subsequent analysis.

Table 1 Variable Descriptions

Field	Description
Player_Age	Age of the player (in years)
Player_Weight	Weight of the athlete (in kg)
Player_Height	Height of the player (in cm)
Previous_Injuries	Whether the player has had prior injuries (Yes: 1, No: 0)
Training_Intensity	Training intensity (range: [0, 1])
Recovery_Time	Days required to recover from an injury (range: 1-6 days)
Likelihood_of_Injury	Probability of injury (1 = Injury, 0 = No injury)

4.2 Data Processing

The data under study may consist of nonlinear datasets, specifically binary classification (nonlinear) problems. The water safety data studied here is a binary classification issue. The following shows the distribution of previous injuries (Previous_Injuries).

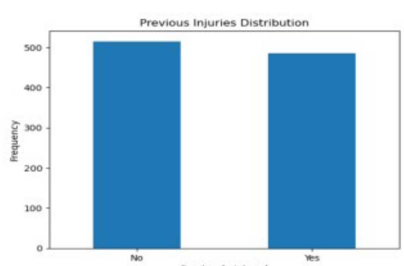

Figure 3-3 Distribution of Previous Injuries

Based on Figure 3-3, the distribution analysis of categorical variables for 1000 athlete samples shows that more than 500 athletes have no previous injuries (non-injury group), while fewer than 500 athletes reported having experienced at least one sports-related injury (injury group).

In this study, we performed data cleaning, missing value handling, feature selection, and standardization by loading a dataset containing athletes' injury history[14]. "Age" and "Training_Intensity" were chosen as the key features for injury risk prediction. The dataset was split into training and test sets to evaluate model performance, and data was standardized to ensure consistency of features. The preprocessed data is shown below:

	intercept	Player_Age	Player_Weight	Player_Height	Previous_Injuries	Training_Intensity	Recover
0	1.0	24	66.251933	175.732429	1	0.457929	
1	1.0	37	70.996271	174.581650	0	0.226522	
2	1.0	32	80.093781	186.329618	0	0.613970	
3	1.0	28	87.473271	175.504240	1	0.252858	
4	1.0	25	84.659220	190.175012	0	0.577632	
5	1.0	38	75.820549	206.631824	1	0.359209	
6	1.0	24	70.126050	177.044588	0	0.823552	
7	1.0	36	79.038206	181.523155	1	0.820696	

Figure 4-1 Preprocessed Data

Additionally, one of the data preprocessing steps in this study is feature selection, which is a critical step in building an efficient machine learning model. The purpose of feature selection is to identify the most influential features from the original dataset that affect the prediction of the target variable (in this case, "Likelihood_of_Injury"). This can not only improve model performance but also reduce the computational resources and time required for model training. The correct choice of features also improves model interpretability, allowing non-technical stakeholders to better understand the reasons behind model predictions.

In the feature selection step of this study, we specifically focused on "Player_Age" and "Training_Intensity"[9]. The choice of these two features is based on the following considerations:

Business Knowledge: Based on prior research and expertise, we know that an athlete's age and training intensity are important factors affecting the likelihood of sports injuries. Younger athletes and high-intensity training are considered key indicators of higher injury risk.

Data Exploration: Initial data exploration analysis showed that "Player_Age" and "Training_Intensity" are significantly correlated with the target variable "Likelihood_of_Injury." This further validated the reasonableness of using these two features as inputs for the prediction model.

Model Simplification: In machine learning, the "Occam's Razor" principle is followed, which suggests simplifying the model as much as possible while achieving similar performance. By focusing on a small number of influential features, we can build an efficient model that is also easier to understand and maintain[10].

Based on the above considerations, these two features were selected as input variables (X) for the subsequent model training and testing, while "Likelihood_of_Injury" was set as the output variable (Y). After feature selection, we further split the dataset into training and test sets, with 80% of the data for training and 20% for testing. This split ratio ensures that the model has enough data for learning while reserving some data for evaluating the model's generalization ability.

4.3 Model Training

During the training of the generalized linear model (GLM) using Python's statsmodels library, especially for the binary classification problem using the logistic regression model, we first created a GLM model instance by passing the dataset and selected model family as parameters. Here, we used sm.families.Binomial() to indicate that we are dealing with a binomial distribution problem, i.e., a typical binary classification issue. Then, by calling the .fit() method, the model starts training on the provided data, automatically optimizing the parameters to find the best model fit.

After training, we obtained an object containing all training results and model parameters with results = glm.fit(). Here, results.params provided the specific values for each model parameter, including the intercept and coefficients for each feature[15]. These parameters helped us understand how each independent variable influences the target variable and whether the influence is positive or negative.

To better understand and present the model parameters, we organized and transformed them into an easily interpretable model equation. By iterating through each parameter in results.params, we constructed a model equation string of the form $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. This equation visually demonstrates the linear relationship of the model, where β_0 represents the intercept, and $\beta_0, \beta_1, \beta_2 ... \beta_n$ correspond to the coefficients of each independent variable. This representation not only facilitated the interpretation of the model but also made the validation and optimization process clearer.

Through these steps, we completed the training and parameter estimation of the logistic regression model, enabling us to directly extract and interpret these parameters, making it easier to demonstrate the relationships between the response variable and the independent variables. This method provides a powerful tool for deeper data analysis and model explanation. The following is the summary result of the generalized linear model:

	coef	std err	z	P>|z|
const	-1.1714	1.273	-0.920	0.358
Player_Age	-0.0011	0.010	-0.111	0.912
Player_Weight	-0.0011	0.006	-0.163	0.871
Player_Height	0.0053	0.006	0.814	0.416
Previous_Injuries	0.1601	0.127	1.256	0.209
Training_Intensity	0.6250	0.224	2.792	0.005
Recovery_Time	-0.0152	0.038	-0.406	0.685

Figure 6-1 Model Summary Results

From the results above, we obtained the final model:

$$log(\frac{p}{1-p}) = \eta = -1.174 - 0.0011 * Player\_Age - 0.0011 * Player\_Weight + 0.0053 * Player\_Height$$
$$+0.1601 * Previous\_Injuries + 0.625 + Training\_Intensity - 0.0152 * Recover\_Time$$

4.4 Model Evaluation

For evaluating binary classification problems, common metrics used for logistic regression include accuracy, recall, precision, and F1 score. These metrics comprehensively reflect the model's performance in various aspects and help in the overall evaluation and selection of the best model. The evaluation results are shown below:

模型预测准确率为： 0.5350
精确度： 0.5714
召回率： 0.4571
F1 分数： 0.5079

Figure 6-2 Evaluation Results

From the evaluation results in Figure 6-2, the Logistic Regression model performed reasonably well in the athlete injury prediction task. The accuracy was 53.50%, meaning the model correctly predicted about half of the samples, slightly better than random guessing, but this performance level may not be sufficient for most applications. Precision was 57.14%, indicating that when the model predicted a sample as positive, there was about a 57% chance it was indeed positive, showing the model's capability in judging positive samples, though it is not very strong. Recall was 45.71%, revealing that the model's ability to capture actual positive samples was limited, with more than half of the positive samples not being correctly identified. The F1 score was 50.79%, reflecting the balance between precision and recall, which was not well balanced. Overall, these performance indicators highlight that the model has some shortcomings in prediction accuracy, positive sample recognition, and false positive handling, which may need to be addressed through data preprocessing, feature engineering, trying different models, and adjusting model parameters to further improve performance. Additionally, it is recommended to use more comprehensive model evaluation metrics to further analyze and improve the model.

5 Conclusion

This study uses the Logistic regression model to analyze and predict athlete injury risks. Through detailed model training and evaluation, several important findings were made. First, the model's performance was moderate, with an accuracy of 53.50%, precision of 57.14%, recall of 45.71%, and an F1 score of 50.79%. These metrics indicate the model's applicability in handling athlete injury prediction tasks. The study shows that age and training intensity are two key factors affecting athlete injury risk, and the model's results further confirm that increases in both age and training intensity lead to higher injury risks.

Additionally, the Logistic regression model has strong interpretability, providing clear contributions of each feature to the injury risk. This offers data support for personalized injury prevention measures. These findings are of significant importance for coaches and medical teams in developing precise training and prevention programs, while also providing a scientific basis for athlete health management and advancing research in sports medicine and sports science.

However, there are some limitations to this study. First, the simulated nature of the dataset may not fully reflect the real-world conditions of athletes, limiting the generalizability of the conclusions. Second, the Logistic regression model has certain limitations when handling complex nonlinear relationships. Future research could explore the use of more advanced machine learning methods, such as deep learning and ensemble learning, to improve prediction performance[11]. Moreover, this study only considered age and training intensity, omitting other potential factors such as sleep quality, nutritional status, and psychological condition that could also influence injury risk. Therefore, future studies should expand the dimensionality of the dataset, attempt more complex models, and incorporate additional influencing factors to enhance prediction accuracy.

Overall, although the model's performance is moderate, this study provides valuable preliminary insights and scientific evidence for the field of sports medicine, particularly with its broad potential applications in personalized injury prevention and health management. Future research can further optimize injury risk assessment methods with more precise data and stronger models, contributing to the continued development of athlete health management and sports medicine

About the Author

Fangyong Wang, Corresponding Author: 1923731117@qq.com

References

[1] Leckey C, van Dyk N, Doherty C, et al. Machine learning approaches to injury risk prediction in sport: a scoping review with evidence synthesis [J]. British Journal of Sports Medicine, 2024.

[2] Martins F, Przednowek K, França C, et al. Predictive modeling of injury risk based on body composition and selected physical fitness tests for elite football players[J]. Journal of Clinical Medicine, 2022, 11(16): 4923.

[3] Luu B C, Wright A L, Haeberle H S, et al. Machine learning outperforms logistic regression analysis to predict next-season NHL player injury: an analysis of 2322 players from 2007 to 2017[J]. Orthopaedic journal of sports medicine, 2020, 8(9): 2325967120953404.

[4] Musat C L, Mereuta C, Nechita A, et al. Diagnostic Applications of AI in Sports: A Comprehensive Review of Injury Risk Prediction Methods[J]. Diagnostics, 2024, 14(22): 2516.

[5] Shaw A, Newman P, Witchalls J, et al. Externally validated machine learning algorithm accurately predicts medial tibial stress syndrome in military trainees: a multicohort study[J]. BMJ Open Sport & Exercise Medicine, 2023, 9(2).

[6] Goggins L, Warren A, Osguthorpe D, et al. Detecting injury risk factors with algorithmic models in elite women's pathway cricket[J]. International journal of sports medicine, 2022, 43(04): 344-349.

[7] Thornton H R, Delaney J A, Duthie G M, et al. Importance of various training-load measures in injury incidence of professional rugby league athletes[J]. International journal of sports physiology and performance, 2017, 12(6): 819-824.

[8] Henriquez M, Sumner J, Faherty M, et al. Machine learning to predict lower extremity musculoskeletal injury risk in student athletes[J]. Frontiers in sports and active living, 2020, 2: 576655.

[9] Briand J, Deguire S, Gaudet S, et al. Monitoring variables influence on random forest models to forecast injuries in short-track speed skating[J]. Frontiers in sports and active living, 2022, 4: 896828.

[10] Farhadian M, Torkaman S, Mojarad F. Random forest algorithm to identify factors associated with sports-related dental injuries in 6 to 13-year-old athlete children in Hamadan, Iran-2018-a cross-sectional study[J]. BMC sports science, medicine and rehabilitation, 2020, 12: 1-9.

[11] Rommers N, Rössler R, Verhagen E, et al. A machine learning approach to assess injury risk in elite youth football players[J]. Medicine and science in sports and exercise, 2020, 52(8): 1745-1751.

[12] Huang Y, Huang S, Wang Y, et al. A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning[J]. Frontiers in physiology, 2022, 13: 937546.

[13] Lövdal S S, Den Hartigh R J R, Azzopardi G. Injury prediction in competitive runners with machine learning[J]. International journal of sports physiology and performance, 2021, 16(10): 1522-1531.

[14] Lu Y, Pareek A, Lavoie-Gagne O Z, et al. Machine learning for predicting lower extremity muscle strain in national basketball association athletes[J]. Orthopaedic Journal of Sports Medicine, 2022, 10(7): 23259671221111742.

[15] Karnuta J M, Luu B C, Haeberle H S, et al. Machine learning outperforms regression analysis to predict next-season Major League Baseball player injuries: epidemiology and validation of 13,982 player-years from performance and injury profile trends, 2000-2017[J]. Orthopaedic journal of sports medicine, 2020, 8(11): 2325967120963046.

[16] Jauhiainen S, Kauppi J P, Leppänen M, et al. New machine learning approach for detection of injury risk factors in young team sport athletes [J]. International journal of sports medicine, 2021, 42(02): 175-182.