

Project

Derrick Robinson

2025-03-01

Topic: Predicting Risk Factors for Athletic Injuries

Introduction:

Athletic injuries, especially in high-impact sports, are a significant concern for both athletes and teams. While advancements in sports science have allowed athletes to recover more quickly from injuries that once were career-ending—such as ACL tears or Achilles tendon ruptures—prevention remains a challenge. The aim of this research is to identify key factors that make athletes more prone to certain types of injuries. These factors could include physical attributes, training regimens, biomechanics, or even external conditions like playing surfaces or environmental factors. By understanding the underlying risk factors, athletic teams can be better equipped to predict, manage, and possibly even prevent injuries before they occur. This could lead to more personalized training programs, injury prevention strategies, and recovery processes tailored to each athlete's individual needs. Such proactive measures would not only enhance player performance but also reduce the overall costs associated with injuries, such as medical bills, lost playing time, and the impact on team success. This is a critical problem for anyone involved in sports—from professional athletic teams and their medical staff to sports equipment manufacturers and fitness trainers. The data science aspect of this problem lies in analyzing complex datasets that include an athlete's medical history, training load, biomechanics, and even genetic predispositions to certain injuries. Machine learning and predictive analytics can be used to model these risk factors and identify patterns that would be difficult to spot through traditional methods.

Research Questions:

1. What physical characteristics (e.g., body composition, flexibility, strength) are associated with a higher risk of injury in specific sports?
2. How do training loads, recovery periods, and rest intervals correlate with the likelihood of injury?
3. Can machine learning models predict the likelihood of an athlete suffering a specific type of injury (e.g., ACL tear, hamstring strain) based on historical data?
4. What role does an athlete's previous injury history play in predicting future injuries?
5. How do external factors such as playing surface (e.g., turf vs. grass), weather conditions, and game intensity influence injury rates?
6. Do certain sports pose a higher risk for specific injuries due to repetitive movements or the nature of the sport (e.g., basketball and ACL injuries, football and concussions)?

Approach

To address this problem, I will first gather comprehensive datasets that include a variety of athlete-related information, such as injury history, physical attributes (e.g., weight, height, muscle mass, flexibility), training data (volume, intensity, rest periods), and biomechanical data (e.g., gait analysis, movement patterns). Additionally, I will incorporate external factors like playing surfaces and weather conditions, which have been shown to influence injury rates. The approach will begin with exploratory data analysis (EDA) to identify key patterns and correlations in the data. I will use statistical techniques to test hypotheses about what factors are most strongly linked to specific types of injuries. Next, I will develop machine learning models to predict the likelihood of injuries based on these factors. These models will be trained using historical injury data

and validated using cross-validation techniques to ensure robustness. Evaluation metrics will help assess the models' effectiveness in predicting injuries accurately. For real-time monitoring, I plan to integrate wearable data from athletes, such as motion tracking and heart rate sensors, to continuously assess an athlete's risk of injury based on their current performance and fatigue levels. These insights could be incorporated into a dashboard for coaches and medical staff to monitor athletes' health and intervene when necessary. Lastly, I will test the feasibility of using these predictive models in a practical setting by collaborating with sports teams to implement the model and track its ability to prevent injuries in real-time. This will provide a practical application for the research and a direct impact on injury prevention strategies. Contribution: This approach will contribute to reducing athletic injuries by providing sports teams with tools to predict which athletes are most at risk and by offering strategies to reduce the likelihood of injuries. It will also allow teams to tailor training and recovery programs based on individual risk profiles, ultimately improving player safety and longevity. Furthermore, by identifying key risk factors, this research could help inform future sports science research, leading to the development of more effective injury prevention strategies and better equipment.

Datasets

1. Injury Prediction Dataset Source: <https://www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset>

```
injury_data <- read.csv("C:/Users/rer12/Downloads/injury_data.csv")
```

```
str(injury_data)
```

```
## 'data.frame': 1000 obs. of 7 variables:
## $ Player_Age : int 24 37 32 28 25 38 24 36 28 28 ...
## $ Player_Weight : num 66.3 71 80.1 87.5 84.7 ...
## $ Player_Height : num 176 175 186 176 190 ...
## $ Previous_Injuries : int 1 0 0 1 0 1 0 1 1 1 ...
## $ Training_Intensity : num 0.458 0.227 0.614 0.253 0.578 ...
## $ Recovery_Time : int 5 6 2 4 1 4 2 3 1 1 ...
## $ Likelihood_of_Injury: int 0 1 1 1 1 0 0 1 1 0 ...
```

- Dataset Overview: This synthetic dataset captures critical attributes such as player demographics, training intensities, recovery times, and previous injury histories. It is designed to assist in predicting the likelihood of injuries based on these factors.
- Data Details:
 - Purpose: To provide a comprehensive set of variables that can be used to predict athletic injuries.
 - Collection Period: The dataset is synthetic and does not represent real-world data collection over a specific period.
 - Variables: Includes player demographics (e.g., age, gender), training data (e.g., intensity, duration), recovery metrics, and injury history.
 - Peculiarities: As a synthetic dataset, it may not capture the full complexity of real-world scenarios. Missing values are not explicitly mentioned.

2. College Sports Injury Detection Source: <https://www.kaggle.com/datasets/ziya07/college-sports-injury-detection>

```
sportsinjury_data <- read.csv("C:/Users/rer12/Downloads/sports_injury_detection_dataset.csv")
```

```
str(sportsinjury_data)
```

```
## 'data.frame': 1000 obs. of 13 variables:
## $ Athlete_ID : chr "A0001" "A0002" "A0003" "A0004" ...
## $ Sport_Type : chr "Basketball" "Tennis" "Football" "Soccer" ...
## $ Session_Date : chr "2024-01-01" "2024-01-02" "2024-01-03" "2024-01-04" ...
```

```

## $ Heart_Rate_BPM      : int 151 114 171 160 120 174 174 123 102 121 ...
## $ Respiratory_Rate_BPM : int 20 20 19 21 22 24 16 21 22 18 ...
## $ Skin_Temperature_C   : num 37.4 36.9 37.4 36.9 36.7 ...
## $ Blood_Oxygen_Level_Percent: num 95.5 98.4 96.2 99.6 99.2 ...
## $ Impact_Force_Newtons  : num 135.3 147.7 81.6 133.5 178.5 ...
## $ Cumulative_Fatigue_Index : num 0.481 0.757 0.585 0.387 0.515 ...
## $ Activity_Type        : chr "Jumping" "Running" "Jogging" "Dribbling" ...
## $ Duration_Minutes     : int 22 81 66 80 51 32 20 53 81 50 ...
## $ Injury_Risk_Score    : num 0.628 0.716 0.659 0.594 0.636 ...
## $ Injury_Occurred      : int 1 1 0 1 1 1 0 1 1 ...

```

- Dataset Overview: This dataset is designed for injury detection and prevention in college sports training. It contains data derived from various biosensors and physiological metrics, including heart rate, skin temperature, blood oxygen levels, and impact force. The dataset also includes activity information and athlete-specific factors, such as fatigue and session duration. The primary goal of this dataset is to predict the likelihood of injury occurring during sports training, allowing for early detection and intervention.

- Data Details:

- Purpose: Injury detection and prevention in college sports training.
- Collection Period: January-September 2024
- Variables: ID, Session Date, sport type, activity type, and training data (e.g., heart rate, respiratory rate)
- Peculiarities: none to report

3. Collegiate Athlete Injury Data Analysis & Forecasting & ML

Source: <https://www.kaggle.com/code/dohaaaaaz/collegiate-athlete-injury-95-accuracy/notebook>

```

collegeinjury_data <- read.csv("C:/Users/rer12/Downloads/collegiate_athlete_injury_dataset.csv")
str(collegeinjury_data)

```

```

## 'data.frame': 200 obs. of 17 variables:
## $ Athlete_ID      : chr "A001" "A002" "A003" "A004" ...
## $ Age             : int 24 21 22 24 20 22 22 24 19 20 ...
## $ Gender          : chr "Female" "Male" "Male" "Female" ...
## $ Height_cm       : int 195 192 163 192 173 180 179 167 166 162 ...
## $ Weight_kg       : int 99 65 83 90 79 75 90 64 91 63 ...
## $ Position         : chr "Center" "Forward" "Guard" "Guard" ...
## $ Training_Intensity : int 2 8 8 1 3 9 5 6 4 2 ...
## $ Training_Hours_Per_Week : int 13 14 8 13 9 14 13 7 19 8 ...
## $ Recovery_Days_Per_Week : int 2 1 2 1 1 3 1 2 2 3 ...
## $ Match_Count_Per_Week : int 3 3 1 1 2 4 4 3 3 3 ...
## $ Rest_Between_Events_Days: int 1 1 3 1 1 1 2 3 3 2 ...
## $ Fatigue_Score    : int 1 4 6 7 2 6 7 2 2 7 ...
## $ Performance_Score : int 99 55 58 82 90 74 97 62 58 62 ...
## $ Team_Contribution_Score : int 58 63 62 74 51 84 56 70 67 52 ...
## $ Load_Balance_Score : int 100 83 100 78 83 99 78 100 80 100 ...
## $ ACL_Risk_Score   : int 4 73 62 51 49 54 84 42 50 35 ...
## $ Injury_Indicator : int 0 0 0 0 0 1 0 0 0 ...

```

- Dataset Overview: This dataset is designed to analyze the impact of complex scheduling algorithms on injury rates and athletic performance in a collegiate sports environment. It provides realistic data for athletes, capturing their demographics, training regimes, schedules, fatigue levels, and injury risks.

- Data Details:

- o Purpose: This project focuses on analyzing a dataset that encompasses various attributes of athletes, including physical characteristics, training intensity, recovery patterns, performance scores, and injury indicators. By conducting a thorough analysis of these factors, we aim to derive actionable insights that can inform training strategies and improve athletic outcomes.
- o Collection Period: Not specified;
- o Variables: Includes player demographics (e.g., age, gender), training data (e.g., intensity, duration), recovery metrics, and injury history.
- o Peculiarities: none to report

R Packages for Predicting Risk Factors for Athletic Injuries:

To carry out my project on predicting athletic injury risk factors, the following R packages will be essential for data manipulation, modeling, visualization, and statistical analysis:

1. Data Manipulation and Cleaning

- dplyr: This package is part of the tidyverse and is useful for data wrangling tasks such as filtering, summarizing, and grouping data.
- tidyr: Also from the tidyverse, this package will help reshape and clean the dataset, including handling missing values and changing the format of data (e.g., pivoting or gathering columns).
- data.table: This package is designed for efficient data manipulation, particularly useful when working with large datasets.

2. Statistical Analysis and Modeling

- caret: A powerful package for building machine learning models. It allows you to perform data preprocessing, model training, and cross-validation, and is useful for training classification algorithms (e.g., decision trees, random forests, logistic regression).
- randomForest: Specifically used for building random forest models, which can be useful for predicting injury risk factors based on various inputs.

3. Data Visualization

- ggplot2: One of the most widely used data visualization packages. It is ideal for creating a variety of plots to explore data and present findings (e.g., bar plots, boxplots, scatterplots, and histograms).
- plotly: Useful for interactive plots, such as dynamic scatterplots or 3D visualizations, which can be helpful for presenting complex models and results interactively.
- corrplot: This package is used to visualize correlation matrices, which can help in identifying relationships between variables (e.g., how training load correlates with injury risk).

4. Model Evaluation and Performance

- ROCR: This package is useful for evaluating model performance in classification tasks. It can help generate ROC curves, precision-recall curves, and compute various performance metrics.
- pROC: For ROC curve analysis and performance metrics (sensitivity, specificity, AUC) evaluation.
- Metrics: A package for various performance metrics like RMSE, MAE, and others for regression models.
- lattice: For advanced plotting techniques and multi-panel plots, useful for comparing multiple models or subsets of data.

Types of Plots and Tables to Illustrate Findings

1. Descriptive Plots

- Histograms: To show the distribution of continuous variables like age, body mass index (BMI), or training load, helping to identify patterns.

- Boxplots: To compare the spread of variables like fatigue, recovery periods, or strength levels across different injury types (e.g., ACL tear vs. hamstring strain).

- Bar Plots: To display categorical data, such as injury occurrence by sport type or by training category.

2. Correlation and Relationship Plots

- Correlation Matrix (Heatmap): To visualize relationships between various factors such as fatigue, age, training load, recovery, and injury type. The corrplot package can help here.

- Pairwise Scatterplots: To explore potential relationships between variables, such as how training intensity and rest time interact with injury risk.

3. Predictive Modeling and Performance Metrics

- ROC Curve: To evaluate the performance of your predictive models in identifying injury risk. This is important for assessing classification models (e.g., logistic regression, random forests).

- Confusion Matrix: A table to summarize the accuracy of classification models by showing true positives, false positives, true negatives, and false negatives.

4. Summary Tables

- Descriptive Statistics Table: A summary table showing the mean, median, standard deviation, and range for key variables (age, training load, previous injuries, etc.) for both injured and non-injured athletes.

- Model Performance Summary: A table showing the performance metrics (e.g., accuracy, precision, recall, AUC) for each predictive model tested (e.g., logistic regression, random forest, XGBoost).

What I still need to learn:

- Advanced machine learning techniques: While I'm familiar with machine learning techniques, predictive modeling for injury risk using sports data could involve complex features such as fatigue, training loads, and recovery cycles.

- What I need to learn:

- o How to effectively apply machine learning models (like decision trees, random forests, and XGBoost) to predict injury risk with imbalanced datasets (more non-injured athletes than injured).

- o How to handle class imbalance in predictive modeling and optimize the models for better recall and precision, not just accuracy.

- o How to use feature selection techniques to identify the most important risk factors for injuries, and how to interpret these features in a sports context.

Data Import and cleaning

To import and clean my data, I Used read.csv() to import the data.

Checked for missing values with sum(is.na(data)).

Data Cleaning:

Removed duplicates with dplyr::distinct().

Converted categorical variables to factors (as.factor()).

Imputed missing values using mice for multivariate imputation.

Non self evident information

Injury risk trends over time (e.g., does injury risk increase mid-season?).

Impact of playing surfaces (e.g., turf vs. grass injury rates).

Fatigue patterns—how does recovery time influence injury risk?

Non-linear interactions—are certain risk factors compounding injury risk?

Ways to view data

By sport—Comparing injury rates in basketball vs. soccer.

By player history—How does previous injury history impact future risk?

By training load—Analyzing athletes with high vs. low training intensity

Slice and dice the data

Stratify by injury type (e.g., ACL injuries vs. hamstring strains).

Group by performance level (e.g., high-performance vs. low-performance athletes).

Summarizing data to answer key questions

Descriptive statistics (summary()) to understand mean training load, recovery times.

Pivot tables to compare injury rates across sports and conditions.

Cross-tabulations (table()) to analyze injury occurrence patterns

Plots and tables to illustrate findings

Histograms—Distribution of training loads.

Boxplots—Injury risk scores by sport.

Heatmaps—Correlation between risk factors.

Scatterplots with trend lines—Training intensity vs. injury risk

Further Questions

How to handle class imbalance (e.g., more non-injured athletes than injured)?

What external factors (weather, surface) have the strongest impact on injury risk?

Can real-time wearable data improve predictions?