Evan Haines & Chris Nicknish

BAIS: 6070 Data Science

Final Project Report

5/11/2025

**Final Project Report**

**Executive Summary**

**1. Can we predict which MLB pitchers are at a higher risk of injury based on historical workload and pitch-level data?**  
Yes. By leveraging historical workload and pitch-level data from 2021–2023, we built models that predict whether a pitcher was likely to get injured in the current subsequent season. Using various classification models, we found that injury risk, to an extent, is meaningfully associated with quantifiable patterns in a pitcher’s prior performance and usage. This demonstrates the potential for data-driven foresight in injury prevention strategies.

**2. Which variables (e.g., innings pitched, pitch type usage, velocity changes, rest periods) are most predictive of future injury risk?**The most predictive indicators included pitch velocity/spin rate, frequency of high-stress pitch types (such as sliders and fastballs), and pickoff attempt frequency. Notably, pitchers with higher velocity and more obscure arm angles (obscure angles and higher spin rate) appeared more injury-prone in the current and following year

**3. How effective are standard classification models in identifying at-risk pitchers?**  
The XGBoost model outperformed the Logistic Regression and decision tree in our initial analysis, achieving the higher Roc-AUC score in predicting whether a pitcher suffered an injury or not. Meanwhile, the decision tree model performed the best when predicting next-season injuries. While the models are not perfect, they demonstrated meaningful lift over random guessing. We noticed more false positives within the decision tree model, which are not necessarily bad, since these athletes could simply be labeled as "high risk." These results suggest that classification models can be a useful early-warning system for future injuries to an extent, despite the rarity and complexity of such events. The rarity and complexity of the injuries did provide some limitations across all of our models and analysis

**4. Can these predictive insights support real-world pitching management?**  
Yes. This analysis provides actionable value for front offices, coaches, and medical staff. Teams can use these models to flag at-risk pitchers before injuries occur, enabling preemptive adjustments to workloads, rest schedules, and pitch selection strategies. For example, monitoring high-velocity pitchers with limited rest, refining pickoff mechanics, and managing pitch mix for those with deceptive arm angles may help reduce injury likelihood in the upcoming season.

**Project Overview**

The goal of this project was to determine whether MLB pitcher injuries can be predicted using a combination of pitch-level and workload data. We first examined in-season injury prediction using historical performance metrics. Based on feedback from our professor, we expanded our approach to target future-season injuries, which has greater practical value for injury prevention and management.

Data Sources

* Statcast data (Baseball Savant): Pitch-level metrics for all pitchers from 2021 to 2024
* Injury records: Scraped ProSportTransactions.com matched to Statcast data via player IDs
* Derived features: Pitcher performance metrics such as innings pitched, rest days, pitch velocity changes, injuries, etc.

**Analysis and Modeling**

Extensive data cleaning and presprocessing took place to ensure that the data was ready for the matching learning models. We removed non-numeric and irrelevant columns like player names and descriptions. Importantly, missing pitch-type data was filled with zeros, assuming pitchers who didn’t throw certain pitches simply had no data for them." We created our binary target variable saw a target variable distribution of 58% to 42%. We also created our second binary target variable for the second, future analysis. Standard scaling was applied to numerical features. We then split the dataset into training, validation, and test sets, stratified by injury status to maintain class balance, and continued to proceed with the analysis.

**1. Within-Season Injury Prediction**

We began by training classification models (Logistic Regression, Decision Tree, XGBoost) to predict whether a pitcher would suffer an injury during the same season, using data available up to each game.

* Target variable: Binary ‘Injury’ column.
* Results: XGBoost achieved the highest ROC-AUC (0.913), identifying features like games played, slider velocity, and pickoff attempts as the most prominent indicators.
* Limitations: Feedback indicated this approach has limited foresight value, as injuries may have already been developing or imminent. We also faced the potential of correlation issues with the games played variable and the determination if an injury occurred in the season. We determined that the high performance of the models was likely due to this.

**2. Future-Season Injury Prediction (Main Focus)**

We revised our outcome variable to indicate whether a pitcher who appeared in season N would be injured in season N+1. We used the same preprocessing steps and the same three model types.

* Target variable: Binary ‘Injured\_Next\_Season’ column.

Models tested: Logistic Regression, Decision Tree, XGBoost

* Results: Decision Tree achieved the highest ROC-AUC (0.643), with all three models performing worse than the previous set. Features like slider velocity and pickoff attempts along with other pitching metrics like fastball spin rate were the most prominent predictors. The lower models scores show how predicting future injuries based on current metrics is a difficult task. However, our best model yielded much better results compared to randomly guessing whether a player would get injured.

**Model Evaluation**

We used various model performance metrics to evaluate their outcomes. Classification reports with accuracy, precision, recall, and f1-scores were our initial evaluation metrics. We visualized these performances with things like ROC-AUC graphs, confusion matrixes, and lift curves. These visualizations allow us to better understand and present the findings from our models. Shown below are the evaluating metrics and visualization from our best models in both our primary (current season injury) and revised (next season injury) analysis respectively:

**Analysis 1**

A screenshot of a computer

AI-generated content may be incorrect.A graph with a line

AI-generated content may be incorrect.A graph of blue rectangular bars

AI-generated content may be incorrect.A blue squares with white text

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.

**Analysis 2**

**A diagram of a network

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.A diagram of a decision tree

AI-generated content may be incorrect.**

**Business Recommendations**

Our analysis supports several data-driven strategies that MLB organizations can implement to reduce pitcher injuries and manage workloads more effectively. These recommendations reflect insights from both our in-season injury prediction model and our more impactful future-season injury forecasting model.

**1. Integrate preseason injury risk assessments into team planning**  
Using models like our Decision Tree classifier, teams can evaluate pitchers’ injury risk before the season begins. This enables staff to flag high-risk players for modified training loads, altered throwing programs, or closer monitoring throughout the season.

**2. Monitor key workload and biomechanical indicators during the season**  
Weekly tracking of pitch counts, average fastball velocity, and pitch type distribution—especially slider usage—can help identify pitchers who are accumulating stress or exhibiting early warning signs of injury. Notably, declines in velocity and overuse of certain pitch types were strong predictors of injury in both models.

**3. Personalize workloads based on risk profiles**  
Pitchers flagged as high-risk—especially those with limited rest between outings, frequent pickoff attempts, or unorthodox arm angles—should have individualized plans. This might include restricting innings, changing pitch selection, or implementing scheduled recovery periods.

Ultimately, injuries are a rare occurrence that can happen out of the blue without any predetermining factors in place. However, our models show that there are some variables that lead to injuries more frequently. While our future models might not be ultra-high-performing, they are still worth considering when we are talking about million-dollar arms within a professional athletic organization.