Ben Gelchinsky
Vaasu Gadiparthi
Ethan Hu

Data Bootcamp Final
Tommy John Surgery Analysis
Presentation

Introduction

As our group began to ideate possible final projects, we realized that each of us shares a love for baseball. The logical successor to this epiphany was to seek out some appropriate metric related to baseball with which we could work to construct a full and thorough final project. We ultimately settled on building a predictive model for Tommy John surgery (TJS) among MLB pitchers. This surgery, also known as Ulnar Collateral Ligament (UCL) reconstruction, is a procedure often needed by modern pitchers, given the extremely high speeds at which baseballs are thrown in current play. These speeds place intense stress on the pitching arm, resulting in relatively common UCL tears in the elbow. The surgery, named after famed pitcher Tommy John (the first to undergo the procedure), replaces or repairs the torn ligament, enabling the pitcher to return to the game. We set out to determine which pitcher characteristics allow the most accurate predictions of the necessity and undertaking of Tommy John surgery. This project uses MLB pitching and batted-ball metrics to build and evaluate a binary classification model predicting the likelihood of Tommy John surgery. Our final model achieved an AUC of 0.67 and an F1-score of 0.66, demonstrating that publicly available performance metrics provide a statistically significant signal for identifying injury risk that effectively exceeds several established benchmarks in the field of sports analytics.

Data Description

Our data comes primarily from two sources. The first is baseballsavant.mlb.com, an official MLB site that contains detailed statistics on all pitchers in the MLB. The data from before 2015 lacks certain categories, so we opted to limit our set to 2015 and onwards. The second is a comprehensive list of every single Tommy John surgery ever performed on an MLB pitcher. The list was compiled by Jon Roegele, a well-known baseball writer and analyst.

To begin our process, we created two separate lists: one of pitchers who had undergone the surgery and another of those who had not. To do this, we filtered the list of all TJS to only those from 2015 and more recent. We then placed all Baseball Savant data from those players into one list and data from pitchers who hadn't undergone the surgery in another. Given that we were aiming to predict a rare occurrence (TJS), the number of pitchers who haven't undergone the surgery is far larger than those who have. If we used all data from Baseball Savant, the dataset might have been highly unbalanced. To combat this, we randomly selected an equal number of pitchers from the non-TJS list to use in the model training. This way, for every one player who received the surgery in our data, there is one pitcher who did not. In total, we selected 153 pitchers of each class (there were 153 Tommy John surgeries since 2015).

The baseball savant site contains a variety of pitching attributes, including plate appearances, plate discipline metrics (strikeout percentage, walk percentage, swing percentage, and whiff percentage), run value indicators (wOBA and expected wOBA), and batted-ball quality measures (sweet spot rate, barrel rate, hard-hit percentage, average best exit velocity, and average hyper exit velocity). We then combined our two lists into one dataframe and modified several characteristics, and accounted for null values.

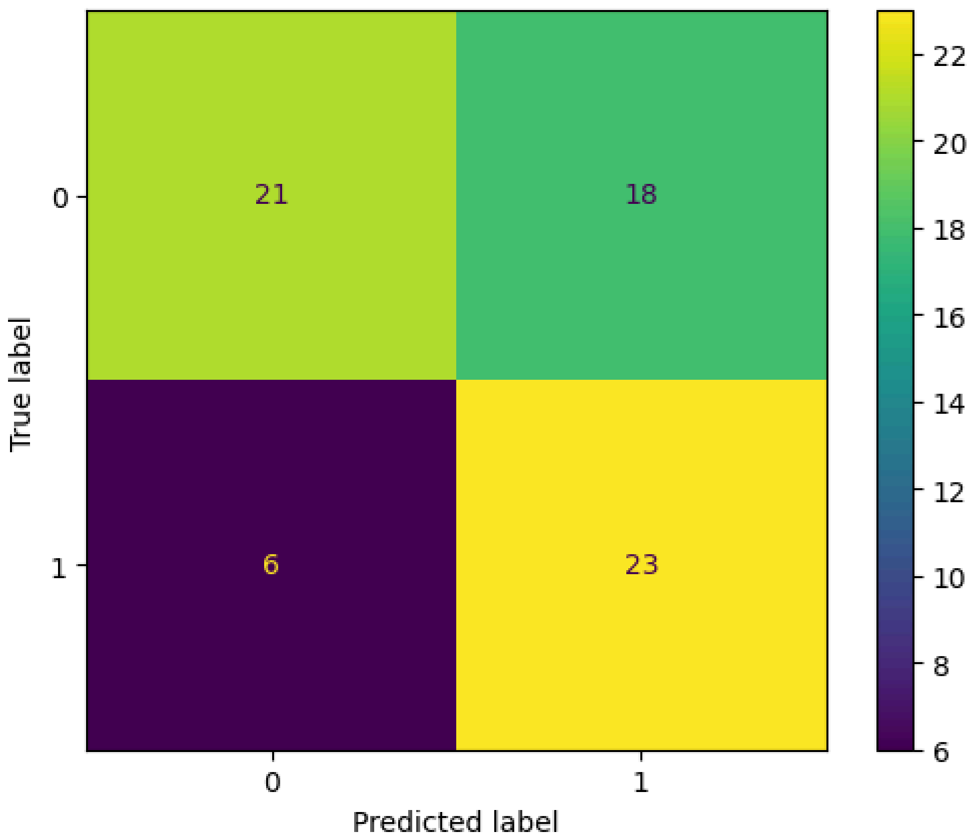
Models and Methods

This preparation complete, we began the process of identifying the significant contributors by employing Scikit-learn's Recursive Feature Elimination (RFE) function. RFE is a wrapper-based feature selection method that iteratively removes less important features based on a fitted predictive model. First, an estimator that provides feature importance measures is trained on the full set of input features. Second, the estimator assigns an importance value to each feature, quantifying its contribution to the model's predictions. Third, a predefined number or proportion of the least important features is eliminated from the feature set. Fourth, the estimator is refit using the reduced feature set, and feature importances are recalculated. This elimination and refitting cycle is repeated until the desired number of features remains. The final output of RFE is a ranked subset of features that retain the strongest predictive power for the target variable while reducing model complexity and potential overfitting. Unfortunately, in our case, the RFE function would reach the iteration limit in Colab. Thus, due to the computational constraints, we excluded pitch-type indicators, as preliminary analysis suggested minimal marginal contribution relative to workload and performance metrics.

For our actual model, we opted to use a binary logistic regression. Our binary outcome (Y) is whether a pitcher undergoes the surgery (1) or not (0). The independent variables (X) are the selected features like player age, games pitched, ERA, etc. The model estimates the probability of a pitcher needing Tommy John surgery based on these features.

Results and Interpretation

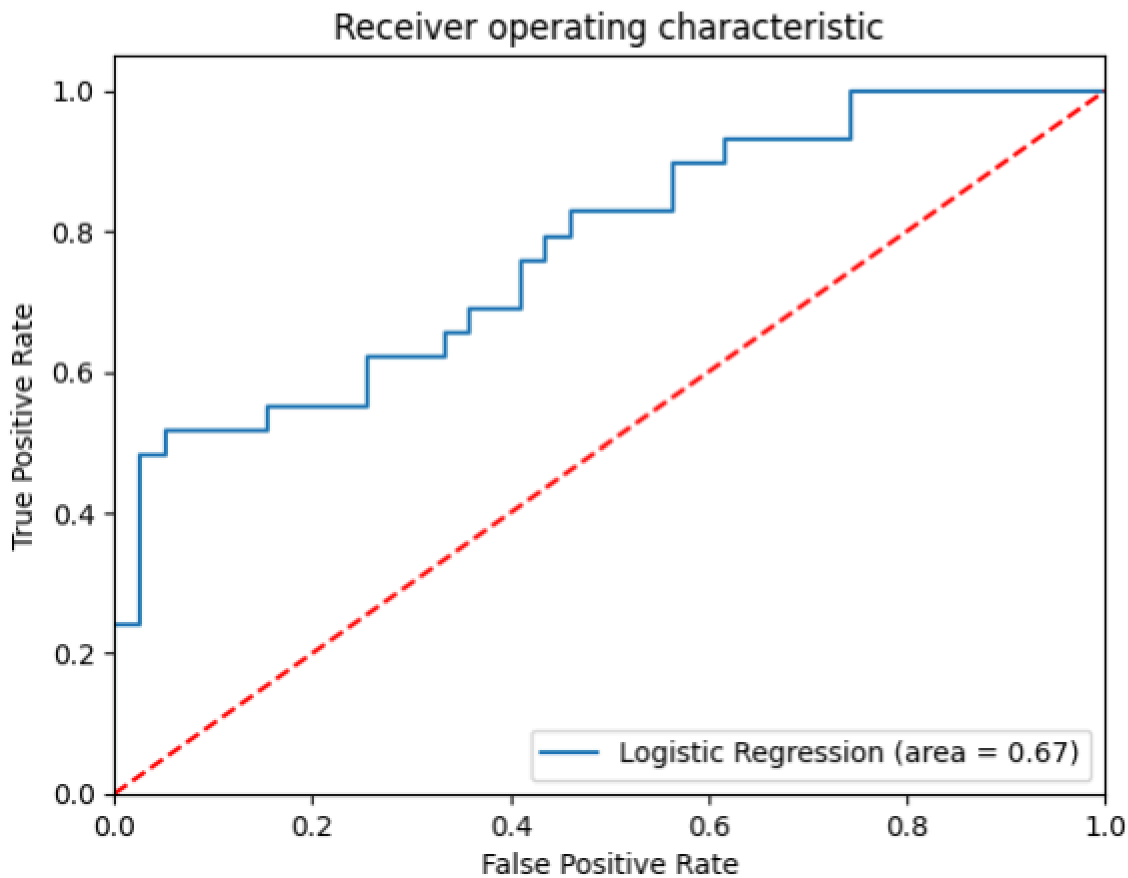
Following the execution of RFE, we found the most important features to be ERA, number of games pitched, K percentage (Strikeout Percentage), and the player's age. After training our logistic regression model, we calculated an F1-score of 0.66. While we tracked overall accuracy, we prioritized the F1-score as our primary metric to ensure the model was effectively identifying surgery cases without being skewed by the high prevalence of healthy pitchers in the league. To investigate the true performance of the model, we constructed a confusion matrix and found these metrics:



- True Negatives (TN): 21 - Correctly predicted non-TJ surgery cases
- False Positives (FP): 6 - Incorrectly predicted TJ surgery cases

- False Negatives (FN): 18 - Incorrectly predicted non-TJ surgery cases
- True Positives (TP): 23 - Correctly predicted TJ surgery cases

We then plotted the Receiver Operating Characteristic (ROC) curve, a graphical representation of the classifier's performance across various thresholds. It plots the True Positive Rate against the False Positive Rate at different threshold settings. The ROC curve helps evaluate the model's ability to distinguish between classes.



A key aspect of the ROC curve is the area under the curve (AUC): The AUC represents the model's overall performance. A perfect classifier has an AUC of 1, while a random classifier has an AUC of 0.5. In our case, the logistic regression model achieved an AUC of 0.67, providing a statistically significant signal that distinguishes injury risk well above random chance, and compares favorably to similar studies, such as a 2022 analysis at Penn State, which yielded a maximum AUC of 0.60 (Rendar, 2022).

Conclusion

For this project, our group set out to develop a predictive model for Tommy John surgery among Major League Baseball pitchers using publicly available performance data. By integrating pitcher-level metrics from Baseball Savant with a comprehensive historical record of Tommy John surgeries, we constructed a balanced case–control dataset and framed the problem as a binary classification task. We selected a logistic regression model while employing Recursive Feature Elimination to identify the most informative predictors and reduce model complexity.

There are several important limitations to our analysis. First, the dataset is observational and therefore cannot establish causal relationships between pitching metrics and injury outcomes. Second, our model does not incorporate biomechanical, medical, or pitch-level tracking data, which are likely critical drivers of elbow injury risk. Third, computational constraints limited the inclusion of certain features, such as pitch-types, which may contain useful predictive information.

In future work, we could extend this analysis in several meaningful ways. Incorporating pitching data tracked across multiple seasons would allow the analysis to model not just whether a pitcher undergoes surgery, but when it occurs, yielding a more realistic approach to injury risk estimation. Additionally, integrating pitch-level Statcast data and biomechanical indicators could improve predictive performance. Overall, this project demonstrates the feasibility of predicting Tommy John surgery risk using publicly available baseball data and provides a foundation for more sophisticated injury modeling efforts in professional sports analytics.

References

Rendar, S. A. (2022). *Predicting Ulnar Collateral Ligament Injury in Rookie Major League Baseball Pitchers* (Honors Thesis). The Pennsylvania State University, Schreyer Honors College.