**Machine Learning and Single-Season Sabermetrics**

**to Identify UCL Injury in MLB Pitchers**

Timothy I. Morales

Department of Statistical Science, Southern Methodist University

STAT 6341: Sports Analytics

Dr. Charles South

December 16, 2020

**Abstract**

**Introduction**

There is money in America’s favorite pastime. Forbes reports that Major League Baseball, MLB, saw a record $10.7 billion in revenue in 2019. In that $10.7 billion includes a $5.1 billion dollar television deal with Fox. The MLB also signed a $1 billion uniform deal with Nike that started in the 2020 season (Brown, 2019). With that huge revenue comes great costs, a $4.7 billion payroll to be exact (Brown, 2019). Fans are paying to watch the MLB, brands are paying to host the MLB, fashion companies want to dress the MLB, and teams want to make the most effective use of their payroll to maximize profits. Since all contracts in the MLB are guaranteed, player injury is at the forefront of mitigating payroll risks, especially injuries associated with long-term performance decreases.

Ulnar collateral ligament, UCL, reconstruction, more commonly known as Tommy John surgery, is notorious in the game of baseball. The sport involving over-hand throwing motions has also been associated with high risks of elbow injury. Over-hand throwing motions can produce extremely high levels of valgus stress in the elbow which eventually leads to instability and increased risk for UCL injury (Chen et al., 2001). Recent research has shown, MLB pitchers are statistically the most at risk group for this UCL injury (Conte et al., 2015). In 2017, Fangraphs a commonly used baseball database, reported that 86.7% of MLB games in the 2017 season included at least one pitcher who had undergone UCL surgery in the past (FANGRAPHS).

Recovery times from UCL reconstruction vary between 12 and 15 months and can span across multiple seasons. Not only does the surgery require long recovery times, recent research has found the surgery linked to performance decreases (Selley et al., 2019). These long absences for recovery as well as long-term performances decreases are concerning for the league, the team, the individual players and the fans. With the afore mentioned billions of dollars going into player payroll each year, understanding and preventing injuries like UCL injuries has huge monetary value.

Since 2010, there have been multiple studies centered around identifying key risk factors for UCL injury. There have also been studies designed around classifying at risk players relative to a control group. In Erickson et. al. (2014), researchers found MLB players who attended high schools in warmer climates to be at a significantly higher risk. Keller et. al. (2015) found that individuals who throw a higher percent of fastballs were more likely to undergo the surgery relative to a control but found in difference in the effect of velocity.

Most similar to the research in this paper is Whiteside et al. (2016) where researchers used support vector machines and naïve bayes models including variables like average spin rates and average velocity to classify individuals against a control group of equal size. Although Whiteside et. al. (2016) was able to classify individuals correctly 75% of the time, this number is an extreme overestimate of the model’s actual predictive power. This 75% is the maximum average five-fold cross validation accuracy from testing over 4000 different models. Each fold was 41 or 42 individuals and had perfect class balance. Therefore, with no hold-out set to assess the optimal model’s true performance, and test folds that are only 42 individuals with perfect class balance, these findings do not stem from a research design that mimics the practical use of a classification model for UCL injury. Roughly 1000 player pitch in the MLB each season and only 2% of those pitchers will undergo UCL injury within the following year. (I hope this paragraph makes Frank Harrell proud. Ironically, I looked the author up and believe this paper got him a job pretty high-up at the Yankees in their biomechanics team).

This paper seeks to address some of the concerns within the previous research as well as build upon prior findings by using a design that more closely emulated the population of MLB pitchers. The classification models in this research also include latest MLB Statcast information. This Statcast information, also referred to as sabermetrics, includes individual pitcher’s spin rates, velocities, movements, and pitch counts across various pitch types. This new Statcast database was released by the MLB in 2015 and is publicly accessible through MLB.com’s Baseball Savant library. Rhinehart (2020) published a thesis comparing the means of these Statcast variables between a control group and a pre-UCL injury group. Although multiple comparisons were made all with an alpha of 0.05, the study found multiple significant difference between the groups and illustrates the potential these variables have in classification.

Through a more holistic and population wide approach, this paper uses single-season traditional counting statistics, the new Statcast database, and personal information including height, weight, and birthplace to build a classification model that correctly predicts individuals who will undergo UCL reconstruction in the next year. This research uses methods including, SMOTE, random forest, neural networks, boosting models, logistic regression and model stacking to correctly classify the entire population of pitchers at risk between the 2015 and 2019 seasons.

**Methods**

***Data Collection***

The analysis involved data from three separate sources, the Lahman database, Jon Roegele’s Tommy John surgery list, and the afore-mentioned MLB.com Baseball Savant database. The Lahman (2020) database is an open-source collection of baseball counting statistics that is easily accessible through RStudio (2020) and is maintained by journalist, Sean Lahman. Single season pitching statistics and single season post-season pitching statistics were taken from the Lahman database for all pitchers who appeared in the 2015 through 2019 seasons. The Tommy John surgery list was obtained from the open-source list kept by Jon Roegele, a former writer for Fangraphs, a well-known baseball statistics database and website. Within the surgery list are all players publicly known to have undergone UCL surgery as well as the surgery date and multiple player IDs.

The Lahman statistics were joined to the Tommy John statistics using baseball reference IDs. The baseball reference IDs for the Tommy John list were obtained from another former Fangraphs, author, Tanner Bell who publicly posts a baseball ID translation sheet on his website, SmartFantasyBaseball. Using the Fangraphs IDs prevalent in both the Tommy John list and the translation list, baseball reference IDs were added to the Tommy John list. The Tommy John list was then merged with the Lahman data using the left\_join function in the dplyr package using baseball reference IDs and year (Whickham).

The Tommy John list was collected on September 23, 2020. For all individuals who underwent surgery in 2020, 2019 season statistics were marked as the season associated with the injury. For all other years, a player underwent UCL surgery before March 31st, that player’s previous season statistics were used and marked as the season associated with the UCL injury. For example, if Jacob deGrom underwent UCL surgery on March 31st, 2019, Jacob deGrom’s 2018 statistics are the flagged as the season where the injury occurred. If Jacob deGrom underwent UCL surgery on April 1st, 2019, Jacob deGrom’s 2019 season was flagged as the season where the UCL injury occurred. The March 31st cutoff for all seasons other than 2020 is used so the data and sabermetrics most closely aligned with the UCL injury are used to fit the model. Using a cutoff of roughly opening day will help amplify the signal, if any within the sabermetric statistics as those who are injured during the season will use that season’s statistics rather than back tracing a full year.

The tradeoff and potential issue of using an early cutoff is running the risk of players standing out due to low amounts of playing time in a season rather than their actual risk factors. This issue was addressed by requiring pitchers to have a minimum of 30 batters faced in order for that pitcher’s individual season to be included in the analysis. The 30 batters faced minimum requirement was adopted from the 10 innings pitched requirement used in Whiteside et. al. (2016). Batters faced was used rather than innings pitched due to the volatility in the duration of an inning especially across pitching roles. The results section also provides addition model prediction results for an additional test set created by following the same methodology and pseudorandom seed but with the cutoff of July 31st. The date of July 31st was chosen as it is well past the mid-way point of the season. The additional results are to help check if the model is solely relying on low playing time as well as see how important recency is for accurate classification.

The MLB.com Statcast variables did not include any ID information other than player names. Player names and year was used to merge the Tommy John and Lahman data frame with the Statcast data frame. Overlapping variables including games pitched and appearance were used to cross reference the player names and assure that the merge was performed properly. This was again done using the *dplyr* package. There were 76 player names that did not overlap between the two datasets, most because of how certain naming conventions like “Jr.” were used. For these players, baseball reference IDs were collected from the baseball reference website itself, again using the aforementioned statistics cross-reference. With the baseball reference ID, the data was merged for 68 out of the remaining 76 players. A total of six players were removed from the data because their identity could not be confirmed. Before enforcing a minimum of 30 batters faced, the data frame consisted of 3645 out of 3651 eligible major league pitchers.

***Data Manipulation***

The fully merged dataset was subset to only include individuals with over 30 batters faced, resulting in 3301 observations with each observation indicating an individual season for an individual pitcher. Of the 3301 observations, 69, or 2.09%, were marked with a flagging variable to denote a season associated with a UCL injury. A postseason flag variable was created denoting if an individual record more an out in the post season. A variable denoting warm birthplace was created with idea and execution from Keller et. al., (2015). The main difference between this variable and the variable used in Keller et. al., was this paper used birthplace rather than high school. All states within the United States denoted as warm in Keller et. al., were denoted as warm in the warm birthplace variable. For those born outside of the United States, individuals born in countries other than Lithuania, Germany and Canada were denoted warm. Variables pitches per game, pitches per out, pitches per batter faced and batters per inning were also feature engineered. All near zero variance variables were dropped other than postseason games started, postseason saves, postseason innings pitched, post season batters faced, and the UCL injury indicator were removed using the *nearZeroVar()* function from the *caret* package (Max Kuhn, 2020).

Missing data was an issue for many of the Statcast and postseason variables. Missing values for variables denoting counts or counting statistics were set to zero. The remaining variables with missing information were missing as a direct result of an individual not qualifying for a certain statistic. These variables where many individuals did not qualify related to individual pitch type information such as spin rate, velocity, velocity difference, and break distance across the different classifications of pitches thrown by MLB pitchers. Since these variables were the main focus of the analysis and all of these variables contained some form of missing data, the variables were binned into quartiles using the *bin( )*  function from the *OneR* package in R (Holger von Jouanne-Diedrich, 2017). The variables were binned into quartiles denoting if the individual during that individual season was in the top 25%, top 25 -50%, top 50-75%, or bottom 25% of all players individual seasons between 2015 – 2019. Along with these four levels the corresponding to quantiles, an additional level denoting the player did not qualify for that statistic was applied to players with missing information for that variable. This binning strategy allowed the research to maintain all individual players while also maintaining some of the variance present in the pitch specific variables. These binned variables and all other factor variables including year, throwing arm, warm birthplace, league, side of plate for batting, made postseason were one-hot encoded using the *step\_dummy( )* function in the *recipes* package in R (Kuhn and Wickham, 2020).

***Data Modelling***

After one hot encoding the data, the resulting data frame was 3301 observations by 337 variables.

**References**

Brown, M. (2019). MLB Sees Record $10.7 Billion In Revenues For 2019. *Forbes. https://www. forbes. com/sites/maurybrown/2019/12/21/mlb-sees-record-107-billion-inrevenues-for- 2019/(accessed May 08, 2020)*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321- 357.

Chen, F. S., Rokito, A. S., & Jobe, F. W. (2001). Medial elbow problems in the overhead-throwing athlete. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, *9*(2), 99-113.

Conte, S. A., Fleisig, G. S., Dines, J. S., Wilk, K. E., Aune, K. T., Patterson-Flynn, N., & ElAttrache, N. (2015). Prevalence of Ulnar Collateral Ligament Surgery in Professional Baseball Players. The American Journal of Sports Medicine, 43(7), 1764–1769. <https://doi.org/10.1177/0363546515580792>

Erickson, B. J., Harris, J. D., Tetreault, M., Bush-Joseph, C., Cohen, M., & Romeo, A. A. (2014). Is Tommy John surgery performed more frequently in Major League Baseball pitchers from warm weather areas?. *Orthopaedic journal of sports medicine*, *2*(10), 2325967114553916.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr:

A Grammar of Data Manipulation. R package version 1.0.2.

<https://CRAN.R-project.org/package=dplyr>

Holger von Jouanne-Diedrich (2017). OneR: One Rule Machine Learning Classification Algorithm with Enhancements. R package version 2.2. https://CRAN.R- project.org/package=OneR

Keller, R. A., Marshall, N. E., Guest, J. M., Okoroha, K. R., Jung, E. K., & Moutzouros, V. (2016). Major League Baseball pitch velocity and pitch type associated with risk of ulnar collateral ligament injury. *Journal of Shoulder and Elbow Surgery*, *25*(4), 671-675.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lahman, S. (2020) Lahman’s Baseball Database, 1871-2019, 2019 version, http://www.seanlahman. com/baseball-archive/statistics/

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86.

<https://CRAN.R-project.org/package=caret>

Max Kuhn and Hadley Wickham (2020). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.14. https://CRAN.R-project.org/package=recipes

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a random forest?. In *International workshop on machine learning and data mining in pattern recognition* (pp. 154-168). Springer, Berlin, Heidelberg.

Rhinehart, E. (2020). Tommy John Surgery: Potential Risk Factors and Causes in Major League Pitchers.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

Selley, R. S., Portney, D. A., Lawton, C. D., Shockley, M. D., Christian, R. A., Saltzman, M. D., & Hsu, W. K. (2019). Advanced baseball metrics indicate significant decline in MLB pitcher value after Tommy John surgery. *Orthopedics*, *42*(6), 349-354.

Whiteside, D., Martini, D. N., Lepley, A. S., Zernicke, R. F., & Goulet, G. C. (2016). Predictors of ulnar collateral ligament reconstruction in Major League Baseball pitchers. *The American journal of sports medicine*, *44*(9), 2202-2209.