THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

COLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

Predicting Ulnar Collateral Ligament Injury in Rookie Major League Baseball Pitchers

SEAN ALEXANDER RENDAR
SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Data Sciences
with honors in Data Sciences

Reviewed and approved* by the following:

Fenglong Ma
Assistant Professor of Information Sciences and Technology
Thesis Supervisor

John Yen
Professor of Information Sciences and Technology
Honors Advisor

* Electronic approvals are on file.

ABSTRACT

The world of data-driven analytics and insight has become one of the most powerful aspects of the corporate realm over the past 20 years. The healthcare industry has followed in suit in using data science for a multitude of purposes. In addition, within the past few years, data science has expanded to the world of professional sports. Most teams, whether it be in the National Football League, National Hockey League, National Basketball Association, or Major League Baseball now employ data scientists or analysts in some capacity. Major League Baseball's use of data analytics is often the most discussed topic. This being the use of sabermetrics, which is the analysis of baseball statistics to answer specific problems. One of the problems still to be answered is an intersection of healthcare and professional sports using machine learning for the purpose of injury prediction. That is, being able to alter athletes' training or workload by discerning injury risks before the injuries occur. Within baseball, pitchers are regarded as the most important players and one of the most common injuries among them is the tearing of the ulnar collateral ligament. Afflicted pitchers are sidelined for the entirety or remainder of the 162-game season. By way of machine learning, modeling has been done using commonly recorded pitching statistics to predict ulnar collateral ligament tears. Injury prediction is a very complex topic in machine learning, like that of fraud detection in the sense of classification bias. The modeling conducted does show promising results, but the findings of this work serve more toward creating a precedence for further research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

A ‘Thank you’ first must go to my thesis advisor, Fenglong Ma, for his time and guidance during the research process and completion of this writing. Without his help, this endeavor would not have been nearly as inspiring, nor as educational.

I would also like to thank the Schreyer Honors College, as well as the College of Information Sciences and Technology for their many academic resources that gave me the foundation for success.

Finally, I want to say thank you to my family. To my parents Gina and James: You have stood by my side through every success and hardship I have faced throughout my life, but especially those of the last four years. Without your endless love and support, I would not be the man I am today. To my sister Mia: Thank you for the advice from your time at Penn State and as a Schreyer scholar. I do not know if I ever would have succeeded as I have without it nor your passion to always be a voice of reassurance. As the last of the four of us to earn a degree from this great school, I am most proud of the fact that I will forever share being a Nittany Lion with those I hold closest to my heart.

Chapter 1

Introduction

In the realm of professional sports, the health and wellbeing of athletes is key to the success of the franchises for which they play. This especially is true with regards to pitchers in Major League Baseball (MLB). A survey poll comprised of MLB general managers, managers (head coaches), and sportswriters found that 60% of the relative importance factor in winning games comes from pitching [1]. That was nearly 40% higher than batting, the second highest factor [1]. They are the first line of defense as they directly control whether their opponent can generate hits and eventually score runs. Additionally, pitchers account for over $1,000,000,000 of MLB salary payrolls, and on average account for over 25% of the payroll per each of the 30 MLB teams [2]. Due to such aspects, losing any rostered pitcher to injury puts teams in drastic competitive and financial hardship, especially if the injury causes a long-term absence from play.

One common injury seen in pitchers is the tearing of the ulnar collateral ligament (UCL). This is a ligament in the elbow which has a massive impact on throwing accuracy, control, and strength [3]. This injury requires a reconstructive procedure commonly known as Tommy John Surgery (Figure 1), wherein a ligament from another area of the body replaces the torn UCL, as seen in figure 2. The namesake for the procedure comes from the pitcher who first underwent UCL reconstruction in 1974 [3]. In a span of 20 years following the first Tommy John Surgery, there were only 12 players who required the procedure [4]. As time went on, the number steadily has grown in that 30 players underwent the procedure between 2011 and 2013 alone and continues to increase [4].

Figure 1. Diagram visualizing a torn UCL [3].

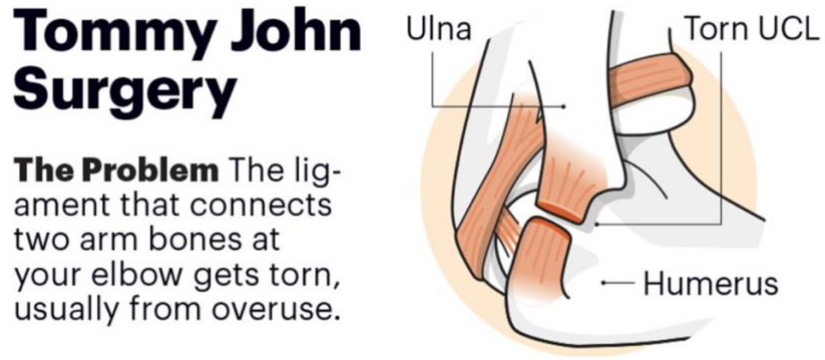
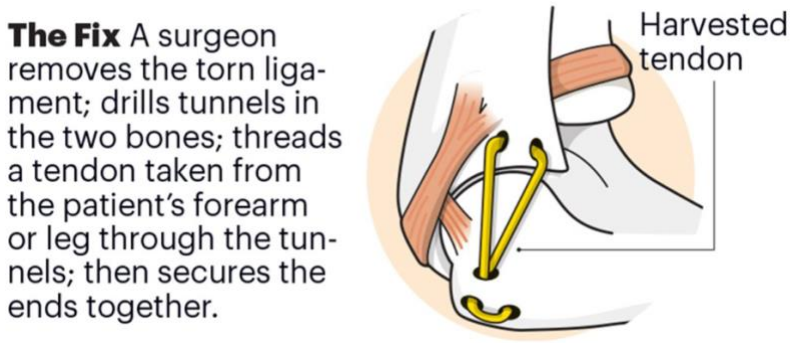


Figure 2. Diagram outlining what Tommy John surgery does [3].



The reason this is such an important topic to research is although the surgery seems minimally invasive, it completely alters the structural composition of a player's elbow. This not only has effects on everyday use of their throwing arm outside of baseball, but it can have permanent negative effects on their throwing form and overall pitching abilities no matter the amount of rehabilitation. The immediate problem is that players who undergo the procedure do not return to play until at least the following season or longer depending on when in the current season the injury occurs. This means general managers face the financial obstacle of having to pay the salary of the injured pitcher, while at the same time finding a replacement. Managers must readjust pitching rotations and game strategies, which could impact a team's record and/or playoff hopes. There are also long-term implications. If a team traded current assets such as draft picks or players of other positions to another team for a pitcher who tears their UCL, then they are

left with a no-win situation. Similarly, if a team focuses their draft picks toward other positions due to the expectation of a pitcher's abilities, they now face such a disadvantage of not having said pitcher or the draft picks to replace them.

The focus of this research is to investigate **if it is possible, by use of machine learning, to predict the requirement for Tommy John Surgery early on in a pitcher's career, more specifically after their rookie season**. By using commonly recorded pitching statistics, specifically from the rookie season of an MLB pitcher's career, it will be reviewed if such statistics can be used to train and test a binary classification prediction of a pitcher requiring Tommy John Surgery. The independent variables will include a multitude of pitching statistics such as total games played, total innings pitched, hits allowed, runs allowed, earned runs average, and others of that nature which will be recorded from pitchers who have undergone Tommy John Surgery in the past. In addition, health statistics such as height and weight will be recorded, as well. The dependent variable will be a binary classification of if a pitcher is at risk of tearing their UCL (positive class) or not (negative class).

The purpose for this is to provide MLB general managers and coaching staffs a more comprehensive ability to investigate pitchers that are early on in their playing careers or possible free agent signings. This allows MLB management the ability to make more educated trading decisions, and adjustments to training and on-season workloads to prevent the long-term downtime and career downturn that follows Tommy John Surgery. The skill and playing ability of these players can directly affect the future success of the franchise they are signed with, significantly. This research will best contribute to the world of sports analytics by providing a novel metric to be observed by MLB management when making roster and contract decisions regarding pitching talent.

Chapter 2

Literature Review

The first article investigated provided a powerful meta-analysis on the intersection of data science, statistical analysis and prediction, healthcare, and sports management [5]. The most important point this article explained was the paradigm shift in the mid-2000s where baseball made a massive shift toward the use of what is now known as Sabermetrics. This is very well outlined in the movie *Moneyball*, where the Oakland Athletics became the first team in major professional sports to operate based on analytics rather than management and coaching intuition. Sabermetrics are the use of statistical analysis to develop advanced decisions based upon in-game MLB events. The writing continued to explain that there exists an underdeveloped area in recording information pertaining to injuries but could exist. Because sports analytics is still predominantly focused on measuring athletic performance for contract negotiations and aspects that could affect how teams perform overall, it is lacking in research for how the same data could be used for predicting and/or preventing injury [5]. The authors then explained that defining a standardized methodology in collecting data for injury risk can only come from a combined effort of data statisticians and sports medicine professionals. In addition to defining such methodology, these fields must also establish it in a way so that is understandable and applicable to be used in real-world settings by coaches, players, and management [5]. This article differed greatly from papers [6], [7], and [8] because it was analytical research on the current stance of sports medicine and data science. The first author, Dr. Marti Casals, is a Spanish epidemiologist with a focus in sports performance analysis research and served as a data analyst for the NBA's Memphis Grizzlies [5]. The second author, Caroline F. Finch, is a research professor at the University of Australia with a focus in sports prevention research [5]. It is believed that this article will be beneficial to this research because it provided insight in how to develop, structure, and justify findings so that they are most useful in sports rather than just data science.

This paper was found to be very informative overall. One aspect they could have discussed more was the way to go about explaining the use of artificial intelligence (AI) and machine learning adequately

so that it is most useful for sports professionals. Given the background of both authors, especially Dr. Casals, he must have techniques that he utilized while being employed by the Grizzlies to best explain to non-data science individuals how his research is justified and practical to NBA players and staff. Likewise, a perspective from sports professionals on their opinions on using data analytics would have also been extremely beneficial to get the other side of who would be using this information.

The next article found proposes a basic workflow for using AI solutions in the realm of sports to prevent participant injury. It also specifically observed 21 soccer players for the purpose of detecting risk factors to prevent injuries as a means of validating their proposed rationale for handling machine learning tasks within sports. Their claim is that by recording information such as emotional well-being, level of activity overall, average exercise load, and a few other advanced statistics, an insightful injury prevention algorithm can be created [6]. By use of Artificial Neural Networks (ANNs), they developed a multilayer classifier to establish non-contact injury risk in said soccer players. They also tested a Decision Tree algorithm (DT) doing the same task. Though they found that DTs are less efficient and more computationally expensive as far as creating output [6]. Sadly, the findings of this paper focused on the efficiency aspects of their models rather than the accuracy of their output. The ANN algorithm had an accuracy of over 90% for predicting injury risk and the DT had an accuracy well over 80% [6]. This writing could be beneficial to this research because it introduced the idea that using data outside of just sports statistics, such as emotional well-being of the athlete. This could provide a more complete picture of the causes of sports injuries overall, although including this type of data is difficult to standardize to be used in machine learning. The most glaring aspect though is that it discussed the use of an ANN deep learning algorithm showing more promising results than regular supervised machine learning.

The main issue with this writing is that its focus was too broad. Its introduction discusses that statistical athlete information can be used to establish a valid machine learning model for injury prevention. Immediately following this, the article shifts focus to discussing using social and economic personal information as an important aspect in using AI in sports injury prevention. Then it dove into a

literature review of other research injury prediction and prevention in a multitude of different sports. Only to shift back and finally discuss the creation of injury prevention models for soccer players. The final portion of the writing was the most powerful area of the work because it introduced modeling types to consider pursuing in this research's application, but due to the article lacking an in-depth discussion on their data or any metrics (besides accuracy) this research will also have a focus of supplying that.

Another article found researched to find advanced machine learning algorithms that could outperform regression analysis in predicting future player availability based on the risk and type of injury for both MLB position players and pitchers. The researchers used player information sourced from online MLB player data as the input values for their work. This included information on previous injuries, number of days they were listed on the MLB disabled list, and other statistics pertaining to player performance [7]. Using python, recurring variables from the different data sources were removed and the data was cleaned of outliers to complete the data preprocessing. Where this research differed greatly from other research is the amount of machine learning models that were trained during the process of creating the prediction, which was 84 total altered models of the following types [7]. These included Linear Regression (a regression analysis model which was compared to the other models used in this research), Random Forest, K-Nearest Neighbors, Naïve Bayes, XGBoost, along with another Ensemble model [7]. These were all tuned using 10-fold cross validation. Accuracy, AUC score, along with F1 were used for metric analysis which found that regression analysis was the 3rd worst model with an accuracy averaging 68% and the best, Ensemble, scored an average accuracy of 70% [7]. They found 7 predictions ranging from full season injuries to as detailed as how many days players would be on the disabled list [7]. They also found that most the significant independent variable was in fact previous injury presence [7]. This research was found to be beneficial because it utilized publicly available MLB statistics from the internet (over 1900 positional player cases and over 1200 pitcher cases) to create a practical prediction algorithm and was able to replicate the findings across the models [7].

It was extremely alarming to see that they tested over 80 models during the research process. They very easily could have scaled down the amount of hyperparameter tuning they worked through by focusing on less models to be included in the research. Their accuracy scores possibly could have been higher if instead of using so many adjusted models, they focused more on data processing to be input into a much smaller number of models. This also would have made hyperparameter tuning much less time consuming when finalizing their findings. What was interesting though is that they included a statistic that from 2000 to 2017, over $435,000,000 was lost due to injury [7]. If they were to expand upon this point by dividing this dollar amount to a per player injury level, it could be useful for MLB franchises in addition to injury prediction.

The final article reviewed undertook using data science machine learning to predict UCL injury in pitchers. Data scientists from the University of Michigan were also trying to find significant statistical variables in predicting the need for Tommy John Surgery. Their purpose in this research was "To identify significant predictors of UCL reconstruction in MLB pitchers" [8]. Using publicly sourced pitcher statistics, 104 MLB pitchers were selected who had undergone UCL reconstruction [8]. A dataset was created based on certain pitching statistics such age, height, mass, innings pitched, pitch count, pitch speed, average days between games, and other quantitative variables that are commonly recorded in the MLB [8]. They then created 5-case samples to better divide the training information which would be the basis for building their models. Likewise, they utilized 5-fold cross validation for hyperparameter tuning to robustly adjust their algorithms. Their main machine learning algorithm in their study was Linear Regression, though they also tested a Naïve Bayes model, along with a Support Vector Machine [8]. From these algorithms, the researchers were able to find significant variables to identify UCL injury with over 68% accuracy in their prediction using Linear Regression, 72% accuracy with Naïve Bayes, and 75% accuracy with their Support Vector Machine [8]. From this, the researchers were able to confidently claim that there are significant independent variables in identifying UCL injury. This research is beneficial because it establishes that a proof of concept similar to what this thesis is hoping to find.

One potential criticism is the size of the dataset they collected from the MLB. They relied on only 104 pitchers as input information for which they based their significant findings. This could potentially have caused bias in their results since, there are over 350 active pitchers in the MLB in any given year. Though it is understood that their goal was not to identify risk of UCL injury but rather identifiers of the injury itself, they could have found a false relation between certain independent variables and injury presence. This is due to the increased amount of positive (has injury) cases they had included in their research versus negative (no injury) cases that are present each year in the MLB. The same variables could be apparent in healthy players and show a more accurate outcome for the prediction, although the significant outcome they discovered would be less as strong. The other issue found is that they only tested using 3 models for classification when there many other algorithms available to be tested. Specifically, K-Nearest Neighbors was not utilized but exists for exactly this type of classification task.

Chapter 3

Project Workflow

To accurately predict the presence of Tommy John Surgery with machine learning, a workflow had to be established to reduce the chance of any data leakage between model training and testing sets, to standardize results between differing models, and to ensure the validity of said results.

Data Collection and Preprocessing. First, was the procurement of the necessary datasets to make the predictions. These datasets consisted of historical pitching statistics of every pitcher's rookie season between 1974 and 2020 (basis of the training set), rookie pitchers of the 2021 season (basis of the testing set), general player dataset to retrieve physical attributes not listed in the pitching datasets, and a dataset of all pitchers who had undergone Tommy John Surgery (target variable). After the data was acquired, the datasets were integrated for the combination of all features and the target classification, categorical variables were all adjusted to a numerical format, and all erroneous values were removed or filled through common tactics found in data analytics.

Feature Selection. Once the data was appropriately cleaned and prepared, feature importance and correlation to the target class were visualized. This was undergone to provide insight into the relationships between the multitude of recorded pitching statistics and the target injury classification. Afterwards, data from past pitchers' statistics underwent feature selection to remove variables found to have insignificant relation to the target class. Once these were identified, the 2021 pitching dataset was reduced to match the variable amount of the historical pitching dataset.

Data Imbalance Issue. After feature reduction, due to a drastic class imbalance of non-injury cases to injury cases, resampling and class weighting had to be undertaken in hopes of increasing the tested models' learning ability for cases showing the requirement of Tommy John surgery. This was done by investigating the use of oversampling of the minority class and undersampling the majority class in supervised learning modeling.

Model Training and Evaluation. Then predictions could then be validly investigated through machine learning. To see if there was an optimal data training set size the training set was proportioned into five different proportions and had all selected modeling types applied to each proportion. During this process, the historical training data was used for hyperparameter tuning cross-validation and model learning. Prediction metrics were then gathered after testing such models on the 2021 rookie pitchers.

Chapter 4

Rookie Pitcher and Tommy John Data

1. Data Collection

Four separate datasets were procured for this thesis. Three of these datasets came from Stathead which is an online-based sports data repository created for the purpose of being a research information platform [9]. The first dataset contained statistical pitching information from the rookie season of each pitcher in the MLB from 1974 to 2020. 1974 was used as the cutoff as it was the first year in which Tommy John Surgery was completed. Within this dataset were 42 features such as team played for, total games played, total innings pitched, hits allowed, runs allowed, and others of that nature. Each row represented a single pitcher out of a total 8,503 cases. The second dataset had the same 42 variables, but only contained pitching statistics for rookie pitchers of the 2021 season. The third dataset contained all MLB players' weights, heights (in inches), their throwing handedness, and their batting handedness since 1974. Stathead did not record the latter four variables in the pitching datasets, so all players in the MLB had to be recorded. These datasets served as the basis for creating the training and testing data to be used in predictive modeling. The training set was constructed from the pre-2021 pitchers, while the testing set was constructed from the 2021 pitchers only.

The fourth dataset (also publicly sourced) provided the target classification variable of this thesis. It contained records of each MLB pitcher who has undergone Tommy John Surgery since the beginning of the procedure [10]. This set not only contained MLB pitcher records, but records on any player of any position. In addition, it also contained records of players at the college level, minor leagues of the MLB, and even international leagues.

2. Data Integration and Preprocessing

Once the data was collected in its raw format of comma-separated values files, they would need reformatted for further use. To begin, the files were uploaded to Google Drive so they could be easily accessed by a Google Colaboratory Python shell. Once this was completed, the data could then be altered using the Pandas data analysis library.

The first step was to reduce the Tommy John Surgery dataset down to only MLB players, and additionally only those players who were pitchers. Next was to assign the binary target class to both the historical rookie pitching dataset and the 2021 rookie set. Though 2021 rookies served as the 'new' prediction in this research, being able to provide it target class gave the research verifiable results. To do this, a new column was created in the pitching datasets by matching the pitchers' names that also occurred in the Tommy John Surgery dataset. If a name matched, the new classification column would record a '1' (injury occurred) for a pitcher and if no match was found a '0' (no injury) was recorded. Then, the dataset of player weights, heights, and handedness was also joined to the pitching datasets so that each would contain these variables as well. Once these steps were completed the data was now successfully integrated from four separate sources down to the two that would be used for modeling.

Furthermore, the datasets would now require any categorical variables to be altered into a numerical format. These variables consisted of the team a pitcher played for, the conference that team was in, their batting handedness, and their throwing handedness. To do this, dummy dictionaries were created. Each team was designated a specific value between 0 to 35 and the pitching datasets where then iterated through to apply this change. The National League conference was identified as a '0' and the American League conference as a '1'. Likewise, both handedness variables were changed to a value between 0 and 2 (3 classes were present due to ambidextrous pitchers who threw and batted with either hand).

The data also suffered from the presence of null, undefined, and infinite values within certain variables. Though it was rare, these values could not be used in the modeling process, so they required data manipulation to adjust them. Null and undefined values were replaced with 0 numerical values, while

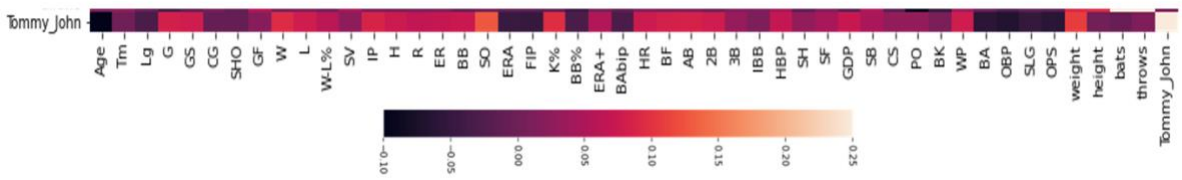
infinite values were replaced by the maximum value found in a column plus a standardized integer value to differentiate them from the true maximum value found in a variable. Once completed the counts of such problem values were validated making sure they no longer existed in the data. A summary of all features is listed in Appendix A under Table 5.

3. Feature Correlation and Importance

To establish interactions between the statistical features used in this research and the target of Tommy John classification, the training data was investigated. The purpose of such procedures is to provide insight toward feature selection and reduction before modeling as unnecessary or unrelated variables could reduce predictive power in machine learning algorithms.

In Figure 3 a heatmap was created to visualize correlations between each of the 46 statistics and the target class. Noticeably, the statistics did not show strong negative or positive correlations to the target variable. The Statistical definition of correlation states that strongly positive correlations are when the coefficient between two variables is at or near +1.0, while strong negative correlations are at or near -1.0. It was found that the range of correlation coefficients in this data was only between the order of -0.1 to just under +0.15. Statistics such as age, earned runs average (ERA), and fielding independent of pitching (FIP) had a negative correlation to the positive injury class. Statistics such as strike outs (SO), weight, and strike out percentage (K%) had a positive correlation.

Figure 3. Correlation heatmap of features to target class.



Though correlation is a beneficial metric in discovering how variables change in coordination with one another, it does not establish a causal relationship. To better understand valid relationships between pitching statistics and Tommy John Surgery, a logistic regression algorithm was created to see how weighted or significant each feature was in deciding the target variable. In logistic regression each feature is assigned a coefficient with an absolute value which describes the strength of a relationship of an independent and dependent variable, as shown in Figure 4. The positive or negative value of a coefficient describes a dependent variable classification of which an independent variable is related. In the case of this research, positive values represent a variable being related to a positive classification of injury, while negative values represent a variable being related a negative classification of injury. From the historical pitching data, it was discovered that variables such earned runs allowed (ER), games played (G), and games started (GS) were of those found to be related to the positive injury classification. Alternatively, variables such as runs allowed (R), age of a pitcher, and inning pitched (IP) were related to the negative injury classification.

Figure 4. Feature importance coefficients.

An issue was identified after reviewing the correlations and regression coefficients. It was noticed that the features R and IP were found in the correlation analysis to show a correlation toward the positive injury class. Conversely, in regression analysis the same two features were found to be significant to the negative no-injury class. This could be due to the presence of intercorrelation between such variables, independent of their individual correlations to the target classification. This intercorrelation would then affect their regression coefficients causing them to show significance toward the wrong class. In an effort to remedy the differing insights, penalty functions of the regression analysis were tested using both ridge and lasso, though both found R and IP to relate to the negative class compared to what was found in correlation analysis.

With this issue being present, and the high chance that this could occur among other features, correlation and regression analysis could not be used for feature selection, but they did serve this research

by providing great data visualization and to show that a more advanced procedure would need to be undergone when selecting features to be used in modeling.

Chapter 5

Model Types

1. Machine Learning

With the combination of computer programming and statistics, machine learning can be defined as the use of algorithms and data to mimic human intelligence by making practical decisions. As the data-age has taken shape, machine learning has been found useful in countless areas such as housing price predictions, disease predictions, text analysis, fraud detection, and many more. Machine learning is based upon the idea that information recorded from past experiences can be trained to a model for the purpose of predicting future outcomes in an accurate manner. The areas of machine learning this thesis focuses on are regular machine learning and deep learning. Machine learning is where models with labeled data are used to predict a condition. As exampled in this research by using structured pitching statistics data to predict a binary target class of requiring Tommy John Surgery or not. Deep learning alternatively does not require structured data such as machine learning. It takes training input and uses to it identify trends and relationships that could go unnoticed by the tactics of regular machine learning. Additionally, deep learning can be used to make decision boundaries based upon a predefined target class (which is its use in this research), or it can be used to make decisions based upon an unknown target class. An example of a use case of deep learning would be to forecast stock market outlooks or to make predictions regarding image data. The predictions made in this research are known as classification. Rather than training on past data to output a continuous value as the target, training is undergone to find two distinct classifications of a target variable. Additionally, the main goal is not only to create prediction models but to create models with accuracy.

The issue in the situation of predicting injury though is that the distribution of classification labels is heavily biased. This means the representation of one class heavily is outnumbered by the representation of another. An example of this in another field would be identifying fraudulent credit card transactions

compared to nonfraudulent charges. This can be a problem because if a trained model does not have sufficient information to train on the lesser represented class, it may very well not gain enough information to make a valid prediction or any prediction at all. To mitigate this issue, data scientists often use tactics such as oversampling or undersampling. Oversampling is the act of creating synthetic data which contains the minority target class. In this research the minority class is a positive identification of requiring Tommy John Surgery. By oversampling the minority class, it can provide models more experience in learning on a situation that does not present itself often compared to other targets. Undersampling is the act of reducing the amount of data which contains the majority target class. In this research the majority class is the negative identification of requiring Tommy John Surgery. By undersampling, the representation of the minority class increases by reducing the amount of data of other classes. The issue with undersampling is that reducing the amount of used data may not be a practical option in some use cases.

2. Models Used

In this research, five models were programed and investigated for the classification of needing Tommy John Surgery. Four of these models were regular machine learning methods and one was a deep learning method. It was felt that by using a variety of model types, the research would be more capable of identifying a valuable prediction through the machine learning workflow.

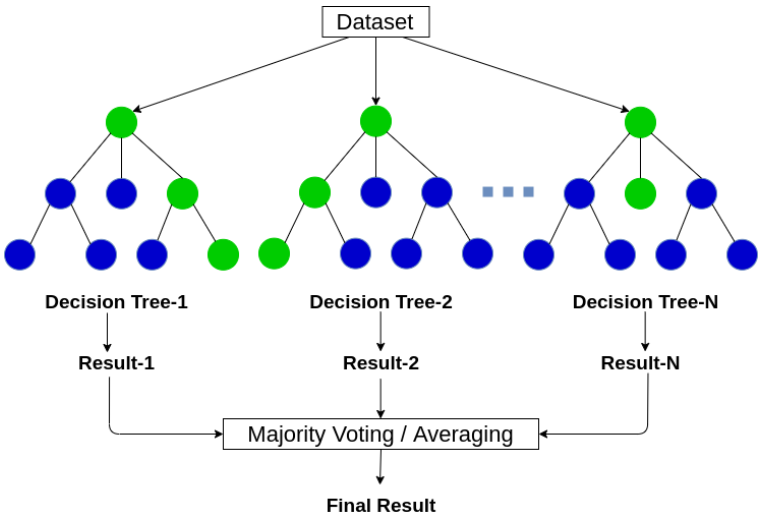
First was a K-nearest neighbors (KNN) algorithm. This is a supervised learning algorithm that works under the methodology that similar outcomes occur under similar conditions. This means that data points (each pitcher's statistics) would be like other data points that have the same outcome. To standardize the idea of similarity, a dimension-space is created depending on the number of variables used in a prediction. A K-value is defined to specify how many 'neighboring' points will be used in classifying the outcome of a data point, which is then used to find either the optimal Euclidean or Manhattan distance

between the K-neighbors [11]. Euclidean distance being the shortest straight segment between two points and Manhattan being the sum of the length of steps in a cartesian plane between points. KNN offers an implementation that would handle a high number of variables and is known for providing proficient classification results when tuned adequately.

The other models used in this research focused on tree-based algorithms. Three tree-based models were utilized for investigation: basic decision tree (DT), Random Forest (RF), and XGBoost (XGB). Tree-based algorithms are especially known for their predictive capabilities in situations of classification, so it was clear that these models would be prevalent in this research. Decision trees operate on the idea of creating nodes that represent features of a given dataset. The branches connecting different non-leaf nodes present conditional if/then statements based upon the feature of the preceding node [12]. The depth or the number of levels of decisions is then adjusted to what provides the most accurate classification of a data point to a target class [12]. The depth of a decision tree is important as it limits the number of conditions that are tested before a label is assigned.

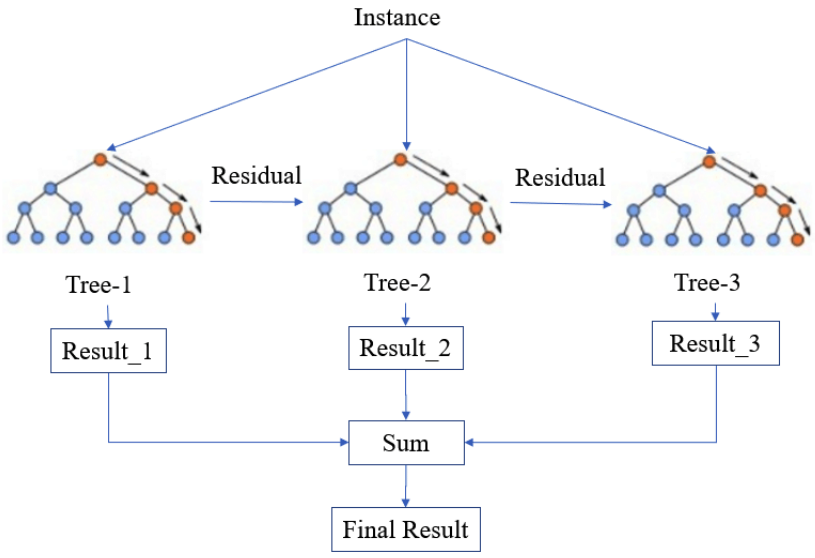
Random Forest builds upon decision trees as it is an ensemble learning algorithm that uses multiple decision trees to create a classification. Bagging, or the simultaneous use of parallel decision trees, is used by Random Forest to reduce the issue of overfitting commonly found in regular decision trees [12]. In addition, Random Forest uses bootstrapping. This tactic randomizes data samples used in creating the differing trees to reduce the change of creating correlated decision metrics within them [12]. Random Forest also enforces feature randomness within each tree to reduce correlation [12]. After each tree in a ‘forest’ is traversed, a classification label is created for a data point by the average decision of said trees. The architecture of Random Forest in shown in Figure 5.

Figure 5. Random Forest Architecture [13].



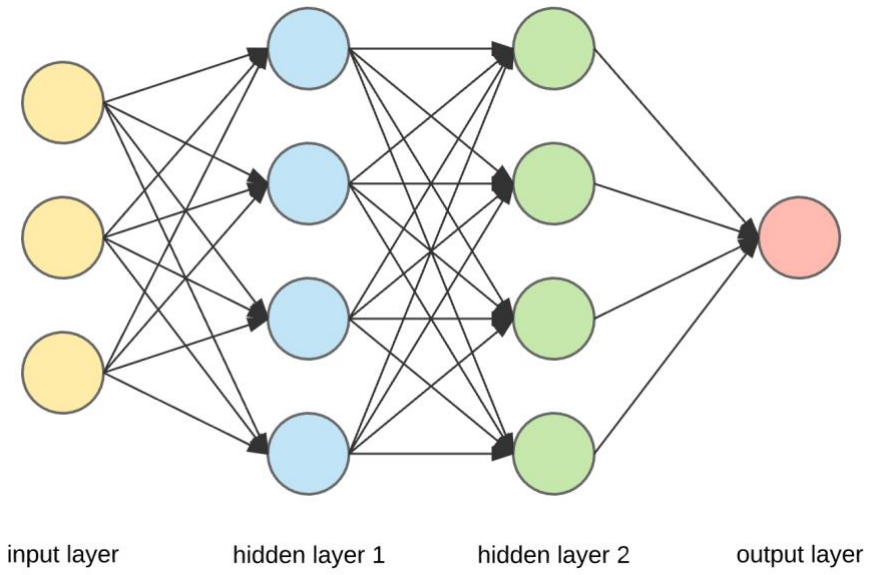
XGBoost, as shown in Figure 6, does not use bagging, rather it uses gradient boosting to achieve a target classification. Gradient boosting can be defined as an ensemble technique where multiple differing iterations of a model are executed to correct errors or residuals of the one ran before it [14]. In the situation of XGBoost, multiple decision trees are created and used to optimize a regression or classification problem. It can provide increased benefit compared to other tree algorithms as it usually has shorter execution times and is known to commonly outperform them in predictive ability [14].

Figure 6. XGBoost Architecture [15].



The final model developed was a deep learning sequential artificial neural network (ANN). ANNs are a deep learning tactic that can solve many types of statistical predictions. Known as universal function approximators, ANNs can model nearly any function, linear or non-linear, at almost any level of complexity [16]. The 'neural' aspect of an ANN comes from the structure of the model. Nodes are constructed into layers and operate similar to how interconnected neurons interact within a brain, with each node operating similarly to a regression algorithm [16]. There are 3 types of layers in an ANN: input layers, hidden layers, and output layers, as shown in Figure 7. Input layers take in data to be trained upon and afterwards to be tested upon. Hidden layers process the data for learning and predicting. Followed by output layers that produce the results. Where deep learning models such as ANNs differentiate themselves from regular machine learning models is that they can find relationships and create decisions based on what is being input, compared to supervised machine learning where models must be taught exactly their purpose and target output [16]. Additionally, most supervised learning models only function on a linear basis while deep learning models are not limited by such a constraint [16].

Figure 7. Structure of a 2-hidden layer ANN. [16]



Chapter 6

Data Manipulation and Model Tuning

1. Feature Selection

Though feature importance was visualized using logistic regression coefficients, it proved not to be the most powerful tactic in selecting the most useful variables during modeling, due to the intercorrelation issues faced. The tactic that was used was SelectKBest feature selection. This feature engineering tool operates on finding the 'K' optimal number of features to train in modeling. The scoring metric used to remove reduce features was 'chi2', which measures the dependence between a target class and a variable (larger values mean more dependence). This scoring metric is designed to be used in SelectKBest feature selection for classification predictions such as this research [17]. When a K-value is specified, the SelectKBest algorithm calculates the 'K' highest 'chi2' scores and records the features which supply them. To investigate the best number of features to be used, K-values from in the range of 5 to 20 were tested and it was found that a K-value of 13 gave the most beneficial results in modeling. This meant that the training and testing datasets were reduced from having 46 pitching statistic features down to 13 features (Table 1) to be used in modeling.

Table 1. 13-Highest Chi2 Scoring Features.

Feature	Chi2 Score	Feature	Chi2 Score
BF	11575.3197	ER	844.203484
AB	10649.6599	R	835.744958
SO	4753.00117	BB	824.999846
IP	2960.61296	GS	673.786328
H	2145.30939	2B	470.681199
ERA+	1148.3872	W	314.662826
G	905.377889		

2. Sampling Manipulation and Cross-Validation Pipeline

To reduce underfitting of a minority positive injury classification, methods of resampling were investigated to establish how to best increase positive class learning ability in the training and predictability in testing. It must be emphasized that to correctly implement resampling strategies, it should only be done on the training dataset. Resampling the testing dataset would alter the distribution of what should be unseen cases for a model to predict. In addition, to properly execute any cross-validation for hyperparameter tuning when resampling, it must work in conjunction with such resampling so that validation sets are without sampling manipulation.

As discussed, the data in this research faced class imbalance. Of the over 8,500 classifications used in training, only ~10% of them represented the positive class for requiring Tommy John Surgery. With a class imbalance of this magnitude there was a high chance that any modeling algorithm would face underfitting the positive classification. The resampling methods tested were undersampling of majority target class (reducing the amount of negative injury classifications), oversampling of the minority target class (duplicating the amount of data of the positive target classification), and Synthetic Minority Oversampling Technique (SMOTE). SMOTE differs from regular oversampling as it does not duplicate data. It uses a KNN algorithm to create new data points between a minority class data point and the closest majority class data point. The resampling tactic that served this research best was oversampling the positive injury class from 10% representation to equal representation to the real majority classification. Multiple oversampling ratios were investigated such as resampling the positive classification presence to 25%, 30%, 50%, and 75%, but oversampling to make the distribution equal was needed to create a sufficient balance between reducing the power of true negative classifications and increasing the power in true positive classifications.

To cross-validate for hyperparameter tuning the machine learning models, a grid-search pipeline was created to work in coordination with the oversampling procedure. The cross-validation style used was repeated stratified 5-fold cross-validation. Both 10-fold and 15-fold pipelines were investigated, but 5-

fold gave the best results in testing. The reason for creating a pipeline to implement cross-validation when oversampling was to ensure that validation sets created by the stratified folds were not affected by oversampling as the training sets for each fold were. Oversampling leaking into the validation sets would drastically alter classification prediction results when testing (usually by overly inflating the minority class predictions). In addition, if the resampling class distributions leaked into validation sets then hyperparameter tuning would cater to the resampling rather than being optimized. The metric used in cross-validation was balanced accuracy (Figure 8) as it is designed to evaluate imbalanced classification problems.

Figure 8. Accuracy compared to balanced accuracy [18].

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

This process would not be undergone for the deep learning model. Resampling would simply redistribute the training set too powerfully, hurting the prediction of the testing set. Cross-validation also is not required. In the case of ANNs, class imbalance is remedied by altering the significance weight of the positive class in relation to the weight of the negative class. This was investigated by setting the positive class weight equal (1:1) to the negative class, twice the negative class weight (2:1), five times (5:1), eight times (8:1), and ten times (10:1). It was found that altering the target class weight to eight times that of the negative class was most sufficient in maximizing performance.

3. Proportioning of Training Data

In addition to oversampling of the positive class and cross-validation, the training data in its entirety was altered using different proportions during training. Each model was investigated using 30%,

40%, 60%, 80%, and 100% of the total training data. Due to the size of the full training set, the resampling that was undergone, and multiple modeling algorithms being used, it would be beneficial to see how differing proportions of training data sizes would affect predicting the test set across differing models. Class distribution was kept equal to the original training set in each proportion and oversampling and cross-validation only occurred after proportioning. This was to provide insight into the stability of the tested models in using a larger sample (80 and 100% of the training data) compared to smaller samples (30%, 40%, and 60% of the training data). It was found that using both the 80% and 100% training dataset proportions provided the most valuable results in testing.

Chapter 7

Modeling Evaluations

1. Evaluation Metrics

The main evaluation metric used in this research was ROC-AUC. ROC-AUC is a commonly used performance metric in binary classification predictions. It measures the true positive rate (correct positive classifications divided by all positive classifications) of a model against its false positive rate (False positive classifications divided by all negative classifications). Meaning it measures the power of a model in predicting the target binary classification of requiring Tommy John Surgery in pitchers. In terms of this metric, a value below the 0.50 line means that a model's predictive ability is not only subpar, but it predicts outcomes worse than randomly selecting a classification target. Values above the 0.50 line mean that a model has some form of predictive ability better thank random chance. A value of 1.00 means that a model perfectly predicted in a testing situation.

For exploratory purposes and a deeper understanding of model evaluations, additional metrics were recorded. These included precision for each predation class, recall of each class, as well as the F1-score of each class. Precision measures the proportion of predicted classifications that were correctly predicted. Recall measures the proportion of actual values that were correctly labelled from a prediction. F1-score works to balance precision and recall, defining a more descriptive metric compared than using accuracy since this is an imbalanced classification problem. As for the investigation for the best predictive ability, ROC-AUC was the metric relied upon for identifying the best models.

2. Machine Learning Prediction Scores

Before overviewing the prediction scores, for clarification purposes the definition of the training and testing datasets and the target classification will be rediscussed. In this research, the training dataset

comprised of rookie pitching statistics from 1974 to 2020. The testing dataset comprised of rookie pitching statistics from 2021 only and from which the ROC-AUC prediction evaluation metric was recorded. The target classification was a binary 2-class variable differentiated by a positive classification of requiring Tommy John Surgery and a negative classification of not requiring Tommy John Surgery.

After fitting the KNN and tree-based algorithms and hyperparameter tuning on the differing proportions of pre-2021 data (training dataset) it was then time to use the 2021 pitching dataset (testing dataset) to test the performance of each model for each proportion. Averaging the ROC-AUC score of running each model five times on each proportion of training data, it was identified that proportioning the training dataset to use 80% and 100% provided the best performance. Interestingly, it was also found that using 80% of the training dataset increased the ROC-AUC score for KNN and Decision Tree models compared to using 100%. For XGBoost and Random Forest, using 100% proved most beneficial. The most powerful machine learning model proved to be XGBoost with a ROC-AUC score of 0.56552 (using 100% of the training data) (Table 2) followed closely by KNN with a ROC-AUC score of 0.56232 (using 80% of the training data) (Table 3). Random Forest was third at 0.55931 (100% of training data) (Table 2). Decision Tree predicted the poorest out of all the models achieving its highest prediction ROC-AUC score of 0.53945, using 80% of the training dataset in learning (Table 3). An overview of the hyperparameter values tested for each model can be seen in Table 6 in Appendix B. These were recorded for the purpose of model replication in future research. Table 7 and Table 8 in Appendix B show all modeling ROC-AUC scores per model within each iteration. In figure 9, graphical representation of the ROC-AUC score can be seen for highest scoring model, XGBoost using 100% of the training data. ROC-AUC curves for the top scoring machine learning models can be seen in Appendix B under figures 11 to 13.

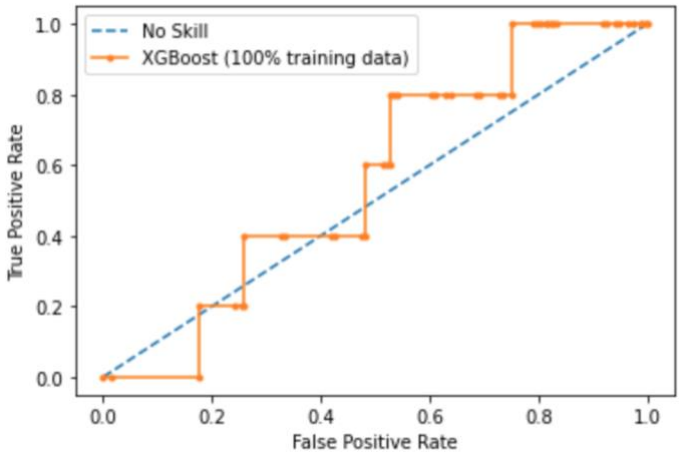
Table 2. ROC-AUC Scores (100% of Training Data).

Model	ROC-AUC
KNN	0.53255
XGBoost	0.56552
Random Forest	0.55931
Decision Tree	0.45945

Table 3. ROC-AUC Scores (80% of Training Data).

Model	ROC-AUC
XGBoost	0.51481
KNN	0.56232
Random Forest	0.51576
Decision Tree	0.53945

Figure 9. ROC-AUC Curve for XGBoost.



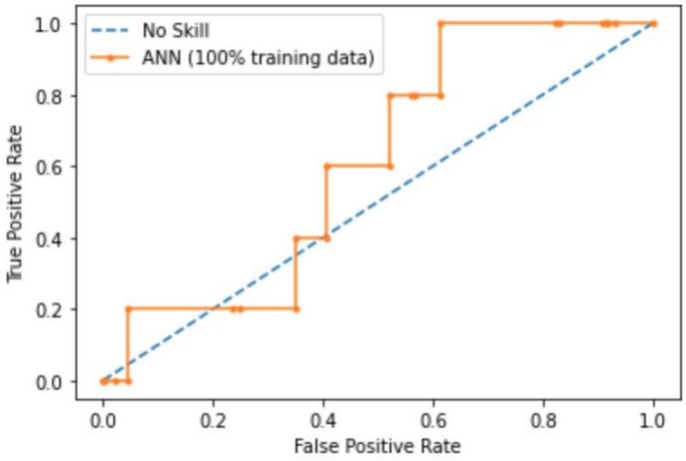
3. Deep Learning Prediction Scores

The ANN was constructed with 2 hidden layers, each with 13 nodes to handle the features selected. Scoring used in training was binary cross-entropy, as is common among using a neural network for classification. Averaging the ROC-AUC score of running each model five times on each proportion of training data was also undergone. Using 100% of the total training dataset proved to provide the highest ROC-AUC with a score of 0.60150, followed closely by the model using 80% of the training set. The deep learning ANN model was able to out-perform all other learning models in ROC-AUC by ~0.04. The layer parameters consisted of the 'relu,' and 'sigmoid' activation functions, and the loss function in which the model was trained using was binary cross-entropy which is commonly used in ANNs for classification. All ROC-AUC scores for the deep learning model can be seen in Table 8 and Table 9 of Appendix B. The ROC-AUC curve for the ANN model using 100% of the training data can be seen in figure 10. In Appendix B, the ROC-AUC curve for the ANN model using 80% of the training data can be seen in figure 14.

Table 4. ANN Performance.

Training Data %	ROC-AUC
100	0.60150
80	0.59426

Figure 10. ROC-AUC Curve for ANN.



Chapter 8

Discussion and Future Work

The research conducted in this thesis oversaw the development of 5 machine learning algorithms to predict the binary classification of MLB rookie pitchers requiring Tommy John Surgery. It was found in the machine learning predications that XGBoost provided the highest ROC-AUC score of 0.56552 using 100% of the available training data. Proportioning of the training data having a noticeable difference in the results of the models was expected, but there was no expectation that limiting the data would improve results for some of the models. Half of the regular machine learning models provided their best ROC-AUC score using only 80% of the training dataset.

The ANN provided the most powerful prediction out of all the models tested with a ROC-AUC score of 0.60150, using 100% of the training dataset. With these findings it must be encouraged to use a sequential ANN in future research regarding this topic. It must be noted that the programming of this model provided the easiest and most robust way of handling imbalanced training data compared to the regular machine learning methods with the use of class weighting compared to the resampling practices of the other models. It also provided the most efficient computational run-times compared to that of machine learning.

In the future, more advanced resampling strategies to greater enhance learning capability should be further researched. In addition, the same should be said for the tuning of class weighting in deep learning predictors for use in injury prediction. Furthermore, non-pitcher MLB players have also shown susceptibility to requiring Tommy John Surgery. If a similar workflow were to be followed, it is wondered if those cases of requiring Tommy John Surgery could be predicted as well. In addition, the exact reasoning as to why the ROC-AUC score increased in some models when using an 80% subset of the total training dataset would be fascinating to investigate.

Though the predictive models created in this research did not provide as powerful of results as it was hoped, at least it is clear that a slight relationship exists between pitching statistics and requiring

Tommy John. Injury prediction is a complex task and requires a massive undertaking of data manipulation through class weighting, resampling, training data proportioning, and feature selection to successfully create usable results. It is hoped that this thesis will act as a foundation for further research on the subject of predicting injury purely based upon player statistics.

Appendix A

Features

Table 5. Statistical Feature Summary.

	mean	std	min	25%	50%	75%	max
Age	24.847348	2.38895605	18	23	25	26	40
Tm	16.5520405	9.54371094	0	9	17	24	35
Lg	0.5402799	0.53702638	0	0	1	1	2
G	17.181583	16.4234487	1	5	11	25	88
GS	3.84405504	6.98606686	0	0	0	4	38
CG	0.19240268	0.90238903	0	0	0	0	24
SHO	0.0571563	0.32249287	0	0	0	0	8
GF	4.35469834	6.50583067	0	0	2	6	68
W	1.91285429	2.80417123	0	0	1	3	20
L	2.11760555	2.78176664	0	0	1	3	19
W-L%	0.32801729	0.3423873	0	0	0.308	0.556	1
SV	0.54192638	2.45171684	0	0	0	0	46
IP	38.1210632	41.9466682	0	8.2	23	53.2	265.2
H	38.81336	41.582268	0	10	24	52	268
R	20.812772	21.4573369	0	6	13	28	130
ER	19.0135246	19.5607897	0	5	12	26	117
BB	16.152064	16.7278237	0	4	10	23	143
SO	27.1759379	30.5404522	0	6	16	38	276
ERA	5.93037399	6.38277963	0	3.38	4.67	6.57	189
FIP	10.1525815	3.16912007	1.96	8.66	9.56	10.725	89.21
K%	15.9265436	7.48764417	0	11.2	15.4	20	100
BB%	10.8210044	5.94610398	0	7.5	10	13	100
ERA+	101.79761	69.6815487	2	66	91	117	1375
BAbip	0.30878255	0.09625932	0	0.263	0.301	0.3435	1
HR	4.32729625	4.9543925	0	1	3	6	36
BF	167.286722	178.736922	1	40	102	233	1120
AB	146.900741	159.303329	0	34	88	205	1019
2B	7.36034341	8.19705488	0	2	5	10	61
3B	0.88886275	1.38216176	0	0	0	1	16
IBB	1.31118429	1.93600894	0	0	1	2	16
HBP	1.4216159	1.91201143	0	0	1	2	19
SH	1.46512995	2.14174851	0	0	1	2	21
SF	1.32376808	1.71634957	0	0	1	2	14
GDP	3.38962719	4.24458315	0	0	2	5	31
SB	2.89474303	4.11606481	0	0	1	4	47
CS	1.26531812	2.02610122	0	0	0	2	20
PO	0.32529695	0.93648119	0	0	0	0	23
BK	0.28060684	0.73175061	0	0	0	0	8
WP	1.60167	2.09641336	0	0	1	2	22
BA	0.27957318	0.08698554	0	0.236	0.271	0.311	1

OBP	0.3626947	0.08798677	0	0.316	0.352	0.397	1
SLG	0.45056153	0.18124368	0	0.356	0.425	0.512	4
OPS	0.81304529	0.2506466	0	0.679	0.778	0.9	5
weight	202.712301	21.802445	143	185	200	215	315
height	74.2178929	2.13013684	64	73	74	76	83
bats	0.77502058	0.48070441	0	0	1	1	2
throws	0.7288016	0.44513319	0	0	1	1	2

Appendix B

Hyperparameters and Results of Machine Learning Models

Table 6. Hyperparameter Ranges of KNN and Tree-Based Models.

Model	Parameter Range
KNN	'n_neighbors': [2, 5, 10, 15, 20, 25, 30], 'weights': ['uniform', 'distance'], 'metric': ['euclidean', 'manhattan']
XGBoost	'min_child_weight': [1, 5, 10], 'gamma': [0.5, 1, 1.5, 2, 5], 'max_depth': [3, 4, 5]
DT	'max_depth': [4, 6, 10, 12, 14, 16, 18, 20]
Random Forest	'n_estimators': [50, 100, 200], 'max_depth': [4, 6, 10, 12, 14, 16, 18, 20]

Table 7. ROC-AUC of each Model Per Iteration (100% of Training Data).

Model	ROC-AUC	Model	ROC-AUC
KNN	0.50784983	DT	0.46416382
KNN	0.5774744	DT	0.46757679
KNN	0.52764505	Random Forest	0.56109215
KNN	0.51945392	Random Forest	0.55085324
KNN	0.53037543	Random Forest	0.5447099
XGBoost	0.55631399	Random Forest	0.578157
XGBoost	0.63276451	Random Forest	0.56177474
XGBoost	0.55017065	ANN	0.60648465
XGBoost	0.53003413	ANN	0.57815701
XGBoost	0.55836177	ANN	0.68600678

DT	0.46279863	ANN	0.51058018
DT	0.44573379	ANN	0.62627983
DT	0.45699659		

Table 8. ROC-AUC of each Model Per Iteration (80% of Training Data).

Model	**ROC-AUC**	**Model**	**ROC-AUC**
KNN	0.59863481	DT	0.61808874
KNN	0.49283276	DT	0.63890785
KNN	0.62662116	Random Forest	0.51604096
KNN	0.5883959	Random Forest	0.55426621
KNN	0.50511945	Random Forest	0.55631399
XGBoost	0.45051195	Random Forest	0.48122867
XGBoost	0.54948805	Random Forest	0.47098976
XGBoost	0.55904437	ANN	0.60443681
XGBoost	0.53720137	ANN	0.57064843
XGBoost	0.4778157	ANN	0.56723547
DT	0.6887372	ANN	0.65426624
DT	0.36723549	ANN	0.57474399
DT	0.38430034		

Figure 11. KNN ROC-AUC Curve.

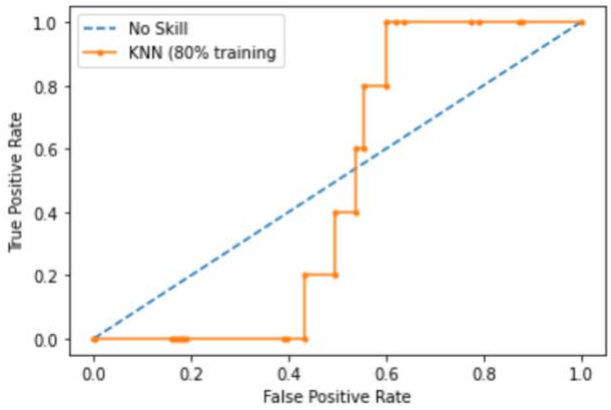


Figure 12. Random Forest ROC-AUC Curve.

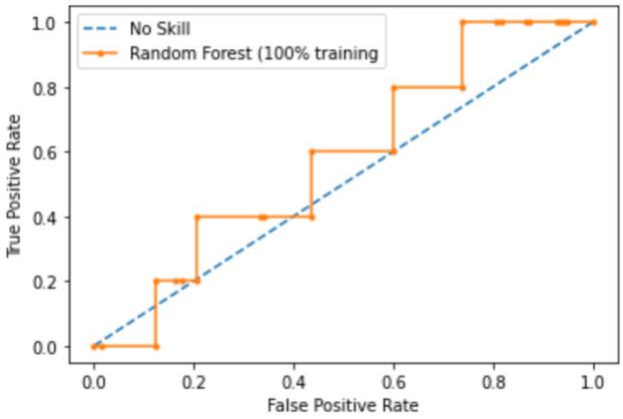


Figure 13. Decision Tree ROC-AUC Curve.

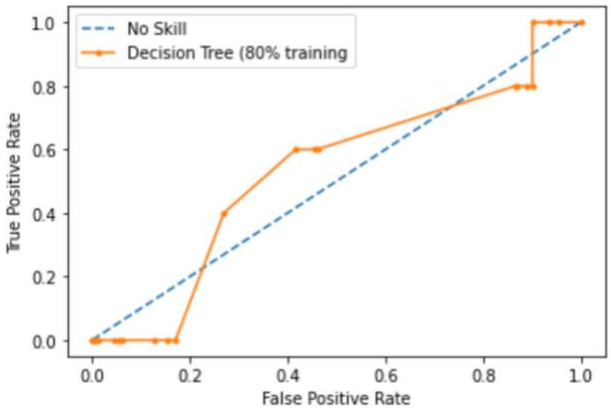
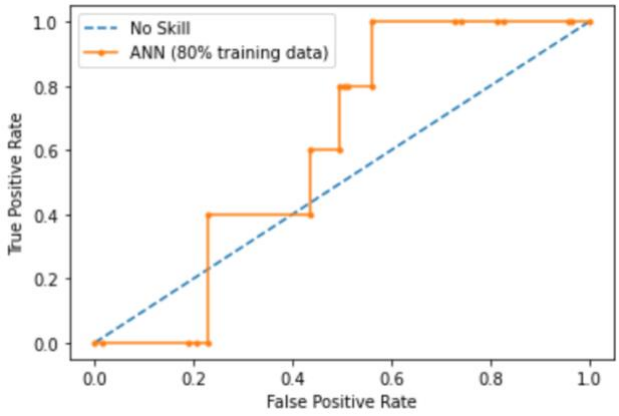


Figure 14. ANN ROC-AUC Curve (80% Training set).



REFERENCES

[1] J. Skipper, Jr. “Is pitching 75% of baseball? Expert Opinions.” *Research Journals Archive*, http://research.sabr.org/journals/is-pitching-75-of-baseball.

[2] “Starting Pitcher Spending.” *Spotrac*, https://www.spotrac.com/mlb/positional/starting-pitcher/.

[3] T. John. “Why Tommy John Is Against the Surgery Named for Him.” *AARP*, https://www.aarp.org/health/conditions-treatments/info-2018/tommy-john-opposes-namesake-surgery.html.

[4] D. White. “The Tommy John Surgery Explosion in the MLB.” *Samford University*. https://www.samford.edu/sports-analytics/fans/2018/The-Tommy-John-Surgery-Explosion-in-the-MLB.

[5] M. Casals and C. F. Finch, “Sports Biostatistician: a critical member of all sports science and medicine teams for injury prevention,” *Injury Prevention*, vol. 23, no. 6, pp. 423–427, 2017.

[6] S. Jauhiainen, J.-P. Kauppi, M. Leppänen, K. Pasanen, J. Parkkari, T. Vasankari, P. Kannus, and S. Äyrämö, “New Machine Learning Approach for Detection of Injury Risk Factors in Young Team Sport Athletes,” *International Journal of Sports Medicine*, vol. 42, no. 02, pp. 175–182, Mar. 2020.

[7] J. M. Karnuta, B. C. Luu, H. S. Haeberle, P. M. Saluan, S. J. Frangiamore, K. L. Stearns, L. D. Farrow, B. U. Nwachukwu, N. N. Verma, E. C. Makhni, M. S. Schickendantz, and P. N. Ramkumar, “Machine Learning Outperforms Regression Analysis to Predict Next-Season Major League Baseball Player Injuries, 2000-2017,” *Orthopaedic Journal of Sports Medicine*, vol. 8, no. 11, 2020.

[8] D. Whiteside, D. N. Martini, A. S. Lepley, R. F. Zernicke, and G. C. Goulet, “Predictors of Ulnar Collateral Ligament Reconstruction in Major League Baseball Pitchers,” *The American Journal of Sports Medicine*, vol. 44, no. 9, pp. 2202–2209, 2016.

[9] “Player Pitching Season & Career Finder.” *Stathead*, https://stathead.com/baseball/player-pitching-season-finder.cgi.

[10] Tommy John Surgery List. @MLBPlayerAnalys, https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit#gid=0

[11] D. Subramanian. “A Simple Introduction to K-Nearest Neighbors Algorithm.” *Towards Data Science*, https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e

[12] S. Yildirim. “Decision Trees and Random Forest.” *Towards Data Science*, https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd

[13] A. Sharma. “Decision Tree vs. Random Forest – Which Algorithm Should You Use?” *Analytics Vidhya, https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/*

[14] J. Brownlee. “A Gentle Introduction to XGBoost for Applied Machine Learning.” *Machine Learning Mastery*, https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[14] “Simplified structure of XGBoost.” *Research Gate*, https://www.researchgate.net/figure/Simplified-structure-of-XGBoost_fig2_348025909

[15] “CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning.” *Analytics Vidhya*, https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/

[16] A. Dertat. “Applied Deep Learning - Part 1: Artificial Neural Networks.” *Towards Data Science*, https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[17] “sklearn.feature_selection.chi2.” *scikit learn*, https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html#sklearn.feature_selection.chi2

[18] “Metrics.” *Uconn*, https://ailab.its.uconn.edu/metrics/#

ACADEMIC VITA

Sean A. Rendar

Education

The Pennsylvania State University — University Park, PA
Bachelor of Science in Data Sciences, Applied Option — 2017-2022
Schreyer Honors College – Pursuing a thesis in Data Sciences — 2021-2022

- Dean's List: Spring 2020 – Fall 2021

Professional Experience

Carnegie Science Center Fab Lab — Pittsburgh, PA
Education Facilitator — May 2021-Present

- Created and facilitated STEM education pertaining to additive technologies

Technical and Administrative Intern — May 2020-May 2021

- Document and advise CSC staff to effectively teach elementary to high school students how to use maker space technologies
- Received the Volunteer Spotlight Award as the most distinguished volunteer

Beaver Area Senior High School — Beaver, PA
Maker Lab Manager — Summer 2019

- Created, planned, and documented Engineering Design curriculum into Robotics and STEM courses for the purpose of improving current state of teaching curriculum and approach
- Maintained lab equipment and coordinated/scheduled usage of the lab (which included 3D printers, vacuum formers, and laser engraver/cutter) for teachers and staff
- Trained both teachers and students on programming and printing process and requirements

Competitive Capabilities International — Beaver, PA
Data Entry and IT Support Intern — Summer 2018

- Supported Senior Vice President, Americas leadership with Executive Committee meeting minutes and follow up
- Supported North America team with expense data entry for client billings and expense reports
- Responsible to provide hardware and software support

Research and Leadership Experience

College of Information Sciences and Technology — University Park, PA
Research Leader — 2020-2022

- Coordinated interviews with a diverse population of Penn State Staff using the Pattee Library 3D Maker Commons; for the purpose of defining how Additive Manufacturing is used in higher education and to aid in improving the Maker Common's functionality
- Formulated research for the development of predictive forecasting of COVID-19 case numbers using Google search trends

Penn State IFC/Panhellenic Dance Marathon (THON) OPPerations — University Park, PA
Supply-Logistics Master — 2021-2022

- Coordinated sanitation supply restocking for the duration of THON weekend

Committee Member — 2018-2022

- Construction and logistics specialist