Modeling Team Success in Major League Baseball Using Machine Learning Techniques

By: Michael Friedrichs

Student #: 501154931

Supervisor: Professor Ceni Babaoglu

December 5, 2022

Ryerson
University

Table of Contents

Abstract

Baseball is a sport that lends itself to accumulating large quantities of useful data, making comparisons between teams' on field performance relatively simple. There are dozens of different statistics that can be used to gauge a team's performance at all three positions; batting, pitching, and fielding. In this paper, an approach based on data mining and machine learning is proposed to identify which of these many statistics is most influential to a team's success, and this information will then be used to predict a team's performance over a full season.

Making reliable projections for baseball teams is crucial for both fans and the organizations themselves who seek to judge a team's underlying talent or predict their future performance. Baseball teams spend tens, even hundreds of millions of dollars on their rosters each year (Spotrac, 2020), so ensuring they are receiving exceptional value in return is of utmost importance. Having an accurate method of projecting success aids greatly when constructing a team, as well as making critical decisions regarding trades or signing free agent players. Fans of the game look to team projections when taking part in sports betting or fantasy baseball, which have exploded in popularity over the last decade (Wikipedia, 2020), to maximize their odds of winning.

Data collected from BaseballReference.com which consists of a team's batting, pitching, and fielding statistics over the last 50 Major League Baseball seasons will be used in this study. With various Python packages, different feature selection methods will be utilized to identify the most significant feature variables, while logistic regression and classification machine learning algorithms will be tested to achieve the most accurate predictive model of a team's success.

Introduction

Predicting the success of sports teams is a challenging task due to the unpredictable nature of sport and the vast number of potential factors that can affect these results. Despite this difficulty, attempting to make these predictions is something that is of interest to many, and this interest has increased with sports data becoming ever more obtainable online, as well as with the growing popularity of online sports betting. Interested parties include bookmakers and sports betting platforms, as well as gamblers who bet on match results or future team success. Sport result prediction is also of great importance to players, team management and performance analysts who use this information to identify which factors are most important to team success and aid in constructing the most competitive team.

The focus of this research paper is on the prediction of a team's success over a full season in Major League Baseball. It will look to answer if machine learning can be used to make an accurate prediction of which MLB teams will finish the season with a winning record, and which teams will be likely to qualify for the playoffs. If so, which machine learning method provides the most favourable results? In addition to these questions, an equally important task is identifying the most pertinent statistics, or features, that will be used by our machine learning models.

As is the case in all team sports, in Major League Baseball, a team must score more points than their opponent in order to win the game. But are offensive statistics more important than defensive statistics when deciding team success? While being able to score a lot of points is

obviously significant, if a team is unable to prevent their opponent from scoring one more point than them, offensive statistics become less relevant. There have been studies conducted on this topic that seem to suggest that both offensive and defensive statistics are both important in baseball (Houser (2005), Fullerton et al., (2014)). However, these previous reports were conducted using simple linear regression on small sample sizes of data and only considered a few statistics that were selected at the writers' discretion. These include common offensive statistics (On Base Percentage, Slugging Percentage, Batting Average) and defensive statistics (Earned Run Average, Errors, Hits Allowed).

It is expected that most of these same statistics will be found significant in the feature selection process in this research project, as they are generally considered among the most important by those who follow the sport. However, using automated feature selection methods on a much more robust dataset may uncover additional statistics not considered in the aforementioned studies. These studies also do not use machine learning to build a predictive model of team success using these statistics, which I suspect can be accomplished with a high level of accuracy.

Major League Baseball consists of 30 teams, split equally into two separate leagues, the American League and the National League. These 15 teams in each league are further split into three divisions, the East, Central, and West. The winningest team in each division will automatically qualify for the playoffs, and the two next best teams will qualify as "wild card" teams, regardless of their division. For example, in 2022, three teams from the American League East made the playoffs, the New York Yankees, Toronto Blue Jays, and Tampa Bay Rays – along with the winners of the American League Central and West divisions. In total, 10 teams

will make the playoffs (five from each league). This was not always the case however, as new rules instituted in 2022 allowed for an expansion of the playoffs from eight teams to ten. And prior to 1994, only the top four teams made the playoffs.

In order to qualify for the playoffs under current rules, a team must have one of the top ten winning percentages out of the 30 teams. While there is no magic number when it comes to qualifying, the general rule of thumb is that a team with at least 89 wins (out of 162) will have a high probability of making the playoffs (Baseball America, 2022). This translates to a winning percentage of at least 0.550, or 55%. For the purpose of this review, win percentage will be converted to a binary target variable in order to perform logistic regression and classification modeling to achieve the desired results.

Literature Review

The following review of literature provides an insight into the historical use of machine learning for predicting sports outcomes. It addresses difficulties that arise in the machine learning process with relation to sports prediction, provides suggestions on how to overcome these difficulties to achieve better results, and describes successful applications of these procedures.

Haghighat et al., (2013) evaluates the use of various data mining and machine learning techniques as an appropriate tool for predicting sports outcomes. After evaluating available literature, the authors highlight the inherent difficulties associated with these techniques. These difficulties arise due to studies usually differing in at least one of the following areas: the particular sport being analyzed, the selected dataset, the input variables, the target variable, and whether individual matches or entire seasons were being considered. The authors conclude that low prediction accuracy highlighted the need for a deeper understanding of the sport to obtain reliable predictions. The variability in the datasets reviewed prevents these researchers from comparing their results with previous studies and leads to unclear development. The authors suggest improving prediction accuracy through the use of machine learning and data mining techniques that have not been used in sports outcomes but have yielded good results in other fields. In addition, the authors suggest including a more robust feature set will help contribute to more accurate predictions.

Expanding on the paper by Haghighat et al., (2013), Bunker & Susnjak (2019) provides a more comprehensive review of surveyed papers. The authors analyzed a total of 31 papers from the

period of 1996 to 2019 to offer insights into the most popular machine learning algorithms used in predicting results in team sports. The article analyzes various characteristics of past studies including the types of algorithms used together with the best performing techniques, the number of features included, and the total number of instances available in each dataset. The article indicates the top three machine learning algorithms in available research are Artificial Neural Networks, Decision Trees, and Ensemble methods including various Boosting algorithms, as well as Random Forests. Bunker & Susnjak conclude that a wide set of candidate algorithms should be used in experimentation in sport result prediction, and that having a rich set of features is more important than having a large number of instances. They claim it is worthwhile experimenting with data-driven methods using various filter-based techniques when selecting these feature variables.

Thabtah et al. (2019) used Naive Bayes, Artificial Neural Network, and Decision Tree models to predict the outcome of NBA Basketball matches. They focused on testing different feature sets in order to find the optimal subset of predictors. The dataset contained 430 NBA finals matches from 1980 to 2017 and 21 features. The authors used three different feature selection methods including Multiple Regression, Correlation Feature Selection, and RIPPER algorithm which is a rule-based classification algorithm often used with imbalanced class distributions. The target variable was a binary variable expressed as a win or a loss. It was found that defensive rebounds was the most important factor influencing match results, as it was selected by all three-feature selection methods. The best performing model in terms of accuracy was trained on a feature set containing eight variables. These were selected from the RIPPER decision rules and subsequently trained using the Decision tree model, which generated 83% accuracy. The authors

suggested that larger datasets and more attributes could be considered going forward to improve performance.

In the study by Jain et al. (2021), results of Indian Premier League cricket matches are predicted using various data mining techniques and machine learning algorithms. These techniques help determine the best methods to generate predictive models that can forecast match results using predefined features in a historical dataset. These features include data based on the performance of teams in past matches, player performance, opposition team information and external factors. Relevant features were selected using Reverse Feature Elimination in conjunction with Random Forest. Machine learning algorithms used include Naive Bayes, Support Vector Machine, K-Nearest Neighbours, and Random Forest and it was investigated which feature subset and classifier together produced best results. The authors conclude that kNN was able to best classify the matches with an accuracy of 70.58%. They also suggest including other feature variables such as field and weather conditions, match timing, and betting odds to improve on future model performance.

Teno et al., (2022) uses data from ten NBA seasons and proposes an approach for predicting game outcomes that is then used for predicting which team will be champion and how far teams will advance in the playoffs. The authors selected a wide range of features including individual game data, team performance from previous years, individual player performance and even salary information. They identified the most relevant features in the model to be the player's performances during the previous season as well as whether a team plays at home. Four models were then used including Logistic Regression, Support Vector Machine, Random Forest and

Multilayer Perceptron. The Random Forest model achieved the best accuracy score of 69.88%. The authors then used this model to simulate 10 NBA seasons 10,000 times and computed frequencies for teams reaching different stages in the playoffs. They were able to show that their approach was equally as successful in picking a Champion as odds makers. The authors suggest that future work can investigate whether this approach can be applied to other sports.

Nguyen et al, (2020) explores the effectiveness of applying Machine Learning to predict the future performance and popularity of NBA players through modelling on players' statistics collected in regular season games. This data was gathered for all NBA seasons between 1982 and 2017, a total of 14,617 observations and 30 feature variables. Player performance was predicted using regression models to determine a player's "Win Share" score, while classification models were used to predict whether a player would make the all-star game. The authors tested a number of classification algorithms including logistic regression, Stochastic gradient descent, support vector machine, random forest, and Naïve Bayes. The models were scored using Accuracy, Precision, Recall, ROC AUC and F1 scores. The article concludes that scoring ability ranks highest in importance for success both for performance and popularity. The article also discusses strategies used to handle imbalanced data, including undersampling, oversampling, and the SMOTE technique. The authors conclude that the under-sampling technique was best suited for their problem compared to over-sampling, resulting in a Recall score of 0.9657 and ROC AUC score of 0.9096 with Random Forest.

It is evident that machine learning algorithms have been used previously to predict outcomes in sports, and in many cases were quite successful. Unfortunately, research into using machine learning to predict the success of a team over an entire season is sparse. Furthermore, there does not appear to be, to my knowledge, any previous work on predicting the per season success of Major League Baseball teams. Most of the available research pertains to predicting individual game outcomes or player performance, and rarely do these relate to baseball. Further research will be required to evaluate the effectiveness of machine learning predicting future team success in not only Major League Baseball, but all major team sports.

The literature reviewed in this report, while not directly related to the research question in this paper, does still help identify some common issues in predicting sports outcomes, and potential methods and strategies to improve model performance. These previous studies can be used as a reference and guide to help achieve the best results possible. In particular, it will be important to have a robust set of feature variables, use different automated methods to select and compare these variables, and utilize a range of machine learning algorithms to achieve the best result. With this information, an opportunity exists to conduct meaningful research in a relatively unexplored domain and to potentially uncover new information that was previously unknown.

Methods

Data Source and Database Creation

The raw data for this study was collected from Baseball-Reference, a readily available and validated sources of Major League Baseball data. The yearly data for each team was downloaded directly from BaseballReference.com in Excel format and then concatenated in Python using the Glob module. The data collected includes all major performance statistics for every team over the last 50 years, averaged over the entire season. Included also are six feature variables outside of player performance that may potentially help strengthen model performance as suggested in the literature review. This results in a dataset consisting of 80 feature variables and 1,396 instances. Performance data includes statistics for hitting (e.g., hits, walks, home runs, on base percentage, slugging percentage), pitching (e.g., strikeouts, walks, earned run average, runs allowed, batters hit), and fielding (e.g., putouts, assists, errors, fielding percentage). Other variables consider the advantage/disadvantage of ballparks played in, the difficulty of schedule, success in games decided by one run, and how many all-star players are on each roster. Brief descriptive statistics and a sample of the dataset, along with a data dictionary and can be viewed at https://github.com/mfriedrichs10/mlbmachinelearning.

Data Processing

As part of the data preparation process, most of the discrete variables were divided by the total number of games played by each respective team each year. This was done to eliminate bias caused by inconsistencies in the number of games played over 50 years of baseball. To illustrate this point, the Minnesota Twins hit 103 home runs in both 1994 and 2011. However, due to the

labor strike in 1994, they hit these 103 home runs in just 113 games as opposed to the full 162

game schedule in 2011. The 1994 team was far more productive in this department in 1994, but

this would go unnoticed by the machine learning algorithms without taking games played into

consideration. Typically, an MLB season will consist of 162 games, and this is the case through

the entire 50 year range of this dataset. Labour strikes in 1972, 1981, 1994 and 1995 however,

led to shortened seasons with 1994 being the most heavily impacted with at most 117 games

being played. The 2020 season only consisted of 60 games due to the Covid-19 pandemic. Other

minor inconsistencies in games played occur over five decades due to tie-breaking scenarios,

where two or more teams have the identical record at the end of the season and one or two extra

games needed to be played to break the tie. Or in some instances, less than 162 games were

played because of postponements due to inclement weather during the season and then these re-

scheduled games being cancelled at the end of the season due to their irrelevancy. These

fluctuations are illustrated in Figure 1.
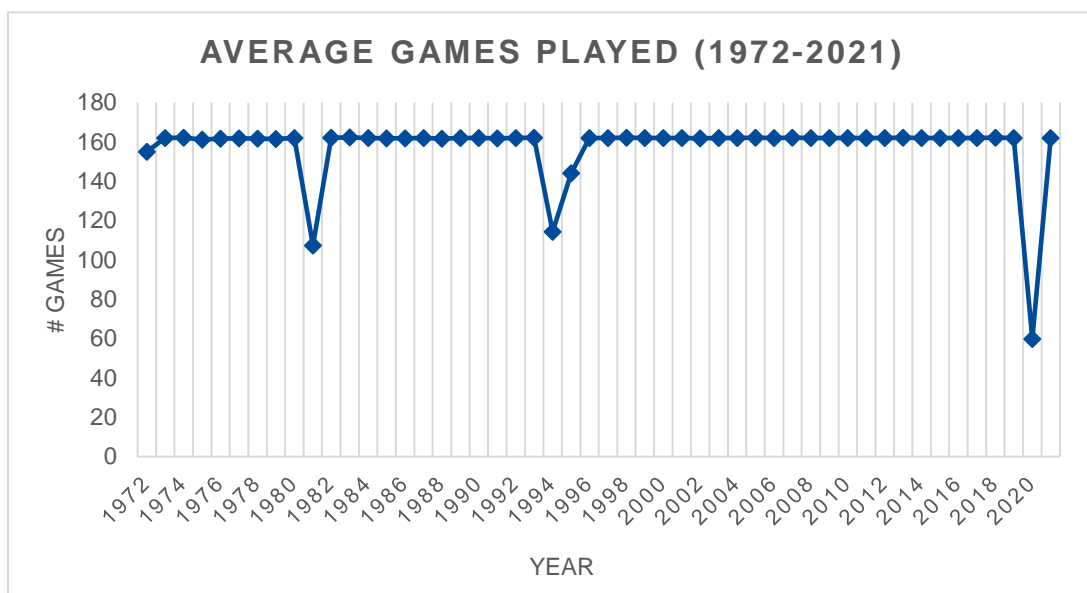

AVERAGE GAMES PLAYED (1972-2021)

Figure 1

Dimensionality Reduction

The original dataset with 79 feature variables is reduced in size by removing features based on relevance, high correlation, and low variance. The Team Name and Year variables are removed as they provide no benefit to model predictions. The variables tSho and cSho (team shutouts and complete game shutouts, respectively) are removed to limit bias, as both features result in a win 100% of the time. The number of games played variable is removed since the majority of the features are averaged by this number as described in the Data Processing section. Finally, due to its extremely low variance across the entire dataset, the Games Started variable is removed since it provides no additional information to the machine learning algorithms.

For the remaining features, a pairwise correlation analysis is conducted to identify highly correlated variables. A correlation coefficient threshold of plus or minus 0.80 is selected to filter out the most highly correlated features. Due to the nature of baseball statistics, there were many of these highly correlated variables in the dataset, resulting in twenty-nine additional variables being removed. The removal of this redundant information should help lower the likelihood of model overfitting and improve the overall training time of the feature selection and machine learning algorithms.

With the irrelevant and redundant variables eliminated, the target variable Win/Loss Percentage is converted to a binary class variable. This is a crucial step as the feature selection and machine learning algorithms selected for this study are suited for classification problems. Fortunately, the Win/Loss percentage is easily converted to a binary variable for both of the modeling scenarios that will be explored. For the prediction of which teams have a winning record, all win percentages below .500 are classified as a losing record (0) and all percentages above .500 are

classified as a winning record (1). To predict which teams can qualify for the playoffs, win percentages below .550 are classified as no playoffs (0) and all percentages above .550 are classified as playoffs (1). For the sake of simplicity, the former dataset will henceforth be referred to as simply Winning Record, and the latter will be referred to as Playoffs.

With dimensionality reduction complete and the target variable converted to a binary variable, the working dataset is finalized. It consists of 44 feature variables and one binary target variable. The target variable for Winning Record is nearly balanced, with 48% of the data being in class 0 and 52% being in class 1. The target variable for Playoffs is more imbalanced, with 74% of the data being in class 0 and 24% being in class 1. See Figure 2.
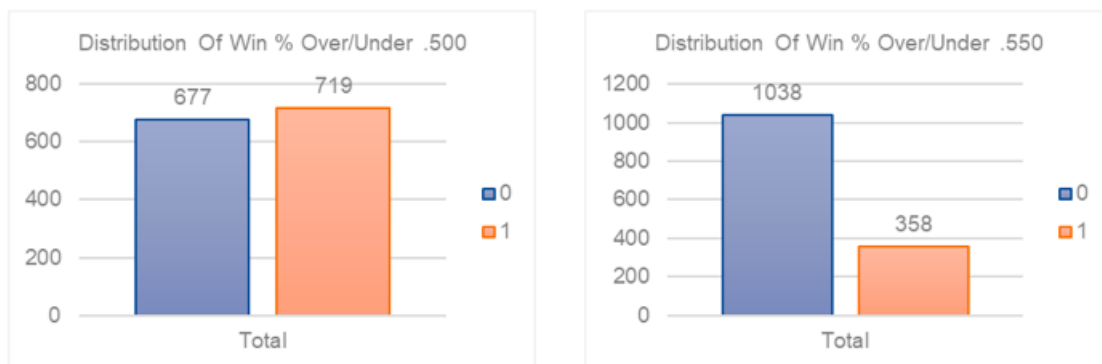

Figure 2

Feature Selection

With dimensionality reduction now complete, feature selection methods can be utilized to identify the most suitable variables for the machine learning models. Mutual Information Gain, ANOVA F-test and Random Forest feature selection techniques will be considered given their ability to handle numerical input data and a categorical target variable.
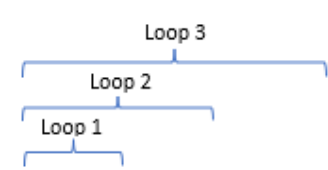
These three feature selection algorithms use different metrics when measuring variable importance, but the end result is an ordered list of feature variables ranked from most significant to least significant. Mutual information calculates the statistical dependence, or the reduction in uncertainty (entropy) between two variables. It is equal to zero when two random variables are independent, and higher values mean higher dependency. ANOVA F-test uses the F-statistic to identify how well each feature differentiates between the two classes of the target variable. The higher the score, the greater the differentiation. Random Forest calculates feature importance by ranking how well they improve the purity of the nodes for each tree. The values are then averaged among all the trees and normalized to 1, resulting in ranked importance scores that add up to 1.

Before identifying the feature ranks, the values of all feature variables are normalized using Min Max Normalization in order to scale the data and preserve the relationships among the original data values. In all three feature selection procedures, the selection is done by examining results on the training set only in an attempt to avoid overfitting. A for-loop is applied to run the train/test split and feature scoring 250 times for each algorithm to achieve more accurate results. After all the feature variables have been scored, the variables (columns) in the original dataset are re-ordered by importance, resulting in three separate datasets - one for each selection algorithm. This process is conducted on both the Winning Record and Playoffs data since the class variable and its distribution is different in both cases.

These three datasets are evaluated to determine what effect the count of feature variables has on the performance metric of the machine learning models. Another for-loop is used to evaluate the performance of each learning algorithm as features are added to the dataset. As an example, for Winning Record classification, loop 1 will calculate the average accuracy achieved by the

machine learning model with only one feature (column). Loop 2 will calculate the average

accuracy with two features (columns) selected. Loop 3 will calculate based on three columns,

and so on until a total of 44 iterations are run, one for each feature variable. An illustration of

this process is shown in Figure 3.

ERA+	OPS+	RA/G	SV	1Run	Under500	BB_P	H_P	R/G
0.173913	0.44	0.477612	0.302094	0.559387	0.402703	0.42918	0.622353	0.339394
0.637681	0.32	0	0.194577	0.390805	0.389189	0.154861	0.181001	0.133333
0.231884	0.56	0.300995	0.265422	0.641762	0.385135	0.42918	0.46492	0.363636
0.275362	0.36	0.161692	0.100396	0.618774	0.271622	0.688073	0.159225	0
0.594203	0.46	0.208955	0.390013	0.39272	0.528378	0.20457	0.482468	0.442424

Figure 3

Using 10-fold cross validation, the accuracy scores for each iteration will be captured, and once

the final loop is complete, the average score for each iteration will be calculated and plotted to

observe the results.

In all three algorithms, model performance increases significantly as additional features are

added, peaking after approximately 5 to 10 features. The Logistic Regression and Random Forest

learning algorithms maintain this level with relatively minor fluctuations until all 44 features

have been processed. Performance for the K Nearest Neighbours algorithm, however, begins to

regress after approximately 20 features have been added in each of the three feature sets. Based

on these results, there is no major observed benefit to keeping all 44 features, and in the case of

the KNN algorithm, it is actually a detriment. Prior to running the machine learning algorithms

in the modeling stage, each dataset will be reduced to the top 20 features determined by the

feature selection algorithms.

Machine Learning Algorithms

The three datasets created during the feature selection process are uploaded. Each dataset contains a different ordering of the feature variables based on their significance. The top 20 features will be selected from each dataset for machine learning, as documented in the previous section. Three different machine learning classification algorithms are now tested – Logistic Regression, K-Nearest Neighbours, and Random Forest. These particular algorithms are selected based on their proficiency with classification problems. A total of nine tests will be conducted, as the data from the three feature selection techniques are coupled with the three classifiers. An illustration of the process can be seen below in Figure 4.
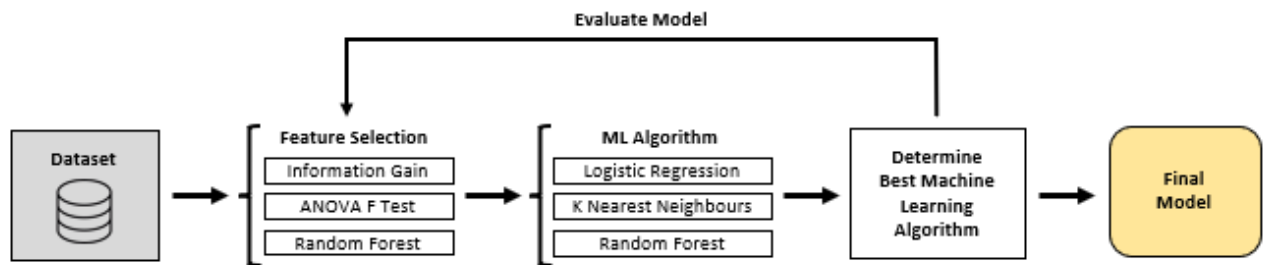

Figure 4

The models are evaluated using confusion matrix results including accuracy, recall, precision, and F1 score. A stratified train-test split is performed on each dataset, with 80% of the data assigned to training set and 20% assigned to test set. Parameter tuning for each machine learning algorithm is performed on the training set using 10-fold cross validation. Since the dataset is relatively small, cross validation was chosen over using separate train, validation, and test sets as this would have resulted in a smaller amount of data being available for training.

The average accuracy and standard deviation is evaluated for each parameter for Winning Record classification. Accuracy is the most appropriate evaluation metric because the positive

and negative classes in this particular dataset are distributed quite evenly and there is no preference towards the positive or negative class. False positives and false negatives are of equal value, the primary objective is to correctly classify as many instances as possible, regardless of their class. When it comes to evaluating the machine learning models for Playoff classification, accuracy is no longer suitable. This is because accuracy does not perform well on imbalanced datasets, which often ends in misleading results. These misleading results occur because if a model does an excellent job of predicting the majority class, but performs poorly on the minority class, it can still achieve a high accuracy score solely due to the class imbalance. In contrast, the F1 score is still able to measure performance objectively when the class balance is skewed. This particular metric is the harmonic mean of the precision and recall values of a model, and the higher the F1 score, the more accurately the model classifies each observation into the correct class. Since there is no preference towards the positive or negative class, as was the case for Winning Record, the macro average of the F1 score is considered most appropriate.

Once the ideal parameters have been identified, the training set is used to fit the model and predictions are then made on the test set. A confusion matrix and classification reports are generated to display the results. Each machine learning algorithm is run three times - once on each set of features selected by the feature selection algorithms. The algorithm parameters are evaluated on each set of features, and adjusted accordingly to achieve the best performance.

Results

Winning Record

Feature Importance

All three feature selection algorithms returned comparable results with regards to feature importance, particularly when ranking the top 50% of the features. Interestingly, the top four features in all three algorithms are the same; ERA+, OPS+, RA/G, and 1Run. This would suggest with reasonable certainty that these are the strongest predictors of team success. These findings, as illustrated in Figure 5, also seem to suggest that there is no clear distinction between batting, pitching, and defense as the most influential predictor of team performance. As was the case in previous studies, all three positions factor into a team's success. One noteworthy result however, was that three of the six external variables that were added to improve the robustness of the feature set ended up ranking in the top ten features. This would seem to indicate that it is not simply individual performance statistics that influence winning. Performing well in 1-run games has been theorized as a positive quality of successful baseball teams, and based on these results, there may in fact be some merit to this. It may signify a team's ability to perform well under pressure. It also makes sense that the #a-tA-S (number of all-stars on a team) and Under500 (number of games against losing opponents) appear among the top ten features. Having highly successful players and a favourable schedule must certainly have an influence on a team's success.
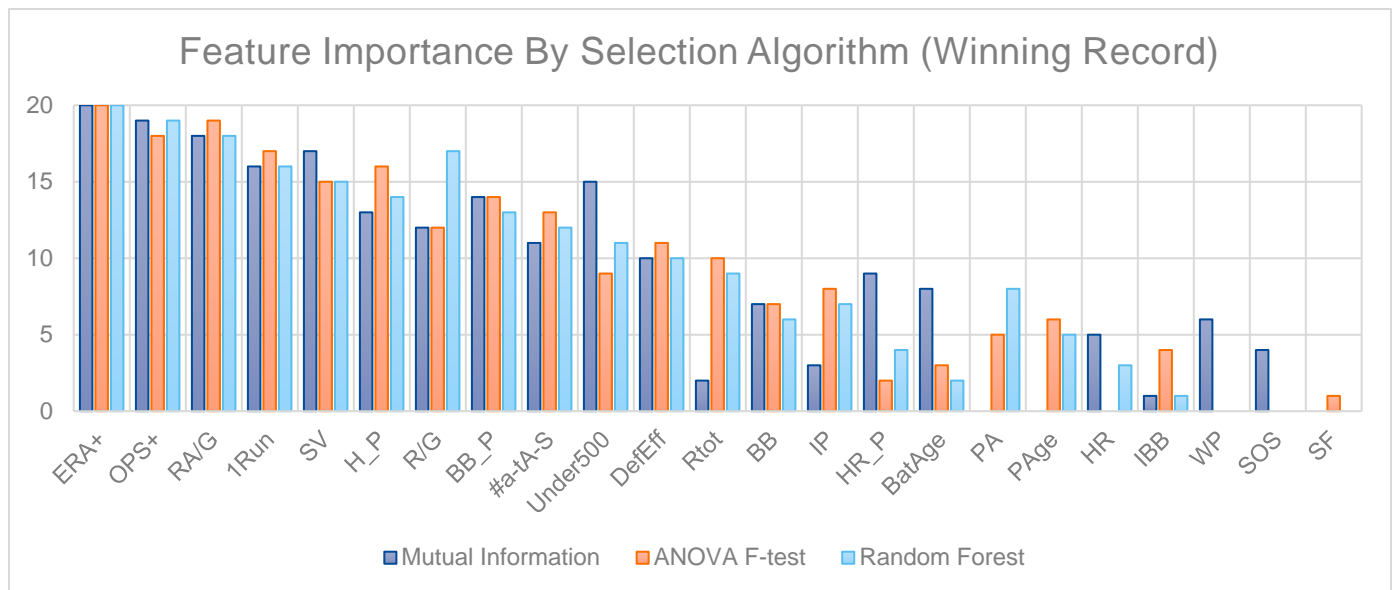
Figure 5 – For illustration purposes, the most important feature is given a rank of 20 and the least important a score of 1

Machine Learning Models

All three machine learning algorithms performed well classifying the data, achieving accuracy scores of over 90%. Through hyperparameter tuning with 10 fold cross validation, a logistic regression model with a L1 penalty term and the liblinear solver resulted in the best accuracy scores results at 94%. It was able to do so consistently with each feature set. Not only did logistic regression produce the best results, the time it took to train and test this model was significantly less than that of the KNN or Random Forest models. This is due to the fact that logistic regression has much lower computational complexity, being that it is a linear model while the other two are nonlinear. The next best performing models were the Random Forest algorithm coupled with the ANOVA F-test feature set, and the K Nearest Neighbours model with the Random Forest feature set. Random Forest was the slowest performing model due to its high computational complexity. Full evaluation results can be viewed in Figure 6.

<u>Winning Record Model Evaluation</u>

Logistic Regression

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.94	0.94	0.94	0.94
ANOVA F-test	0.94	0.94	0.94	0.94
Random Forest	0.94	0.94	0.94	0.94

K Nearest Neighbors

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.90	0.90	0.90	0.90
ANOVA F-test	0.91	0.91	0.91	0.91
Random Forest	0.92	0.92	0.92	0.92

Random Forest

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.91	0.91	0.91	0.91
ANOVA F-test	0.92	0.92	0.92	0.92
Random Forest	0.91	0.91	0.91	0.91

Figure 6

Playoffs

Feature Importance

The top ten features in the Playoffs dataset are nearly identical to those in the Winning Record dataset, with ERA+, OPS+, RA/G, and 1Run again ranking as the top four most important. This would seemingly reinforce the notion that they are the best predictors of team success, and this was not altered by the imbalanced class distribution for the target variable. Notably, the variability of the bottom ten features increased with this imbalanced dataset. Figure 7 illustrates the feature importance rankings by feature selection algorithm.
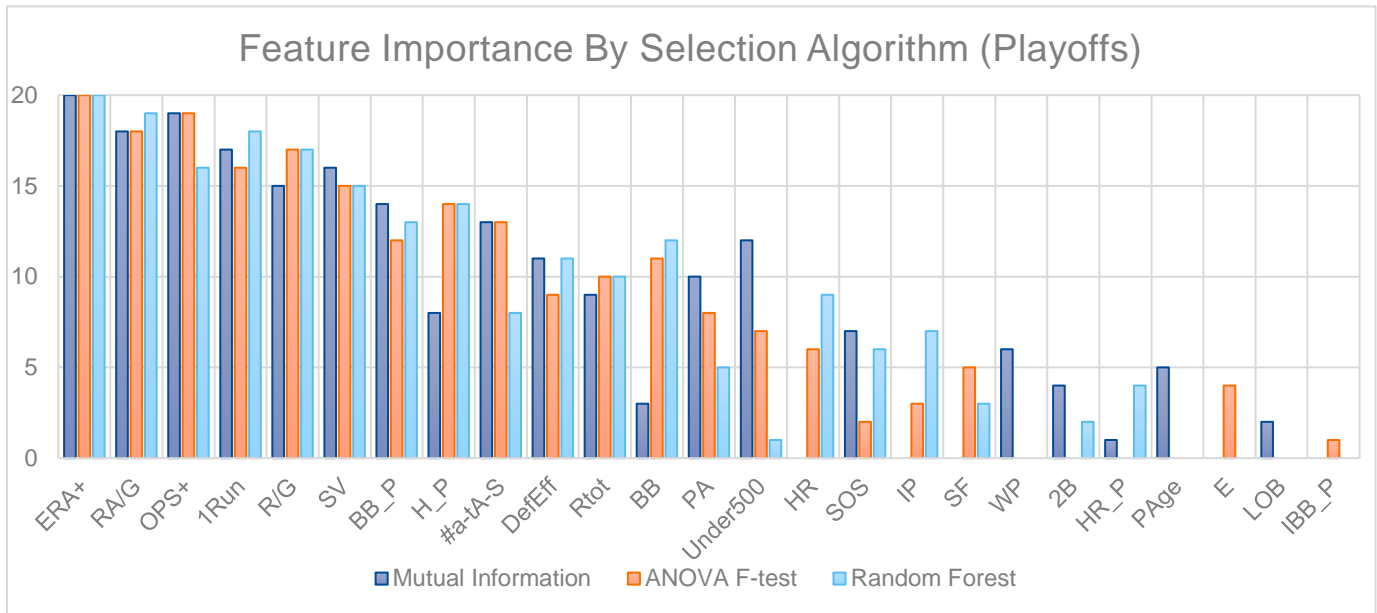
Figure 7 - For illustration purposes, the most important feature is given a rank of 20 and the least important a score of 1

Machine Learning Models

The Logistic Regression model performance remained strong when classifying the imbalanced Playoffs dataset, while the K Nearest Neighbours and Random Forest experienced a drop in performance. The ability of the KNN model to accurately classify the data was impacted the most by the introduction of an imbalanced target class. As was the case with the Winning Record classification, logistic regression performed the best of the three models by achieving the highest macro average F1 score of 93% and it did so consistently with each feature set. With this imbalanced dataset however, the cross-validation results identified the saga solver along with the L1 penalty term to be the best combination of hyperparameters. Logistic Regression also remains the fastest model for training and testing speed. The Random Forest algorithm was the next best performer and the KNN algorithm delivered the least favorable results.

<u>Playoffs Model Evaluation</u>

Logistic Regression

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.95	0.93	0.93	0.93
ANOVA F-test	0.95	0.93	0.93	0.93
Random Forest	0.94	0.93	0.92	0.93

K Nearest Neighbors

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.90	0.82	0.90	0.85
ANOVA F-test	0.90	0.85	0.89	0.87
Random Forest	0.89	0.81	0.90	0.84

Random Forest

Feature Selection	Accuracy	Recall	Precision	F1
Mutual Info Gain	0.92	0.87	0.91	0.89
ANOVA F-test	0.91	0.87	0.90	0.88
Random Forest	0.93	0.88	0.92	0.90

Figure 8

Conclusion

This study affirms the potential of machine learning in the prediction of Major League Baseball team performance. Logistic regression outperformed the K Nearest Neighbour and Random Forest models both in predictive performance and train and test speed – and did so on both the balanced and imbalanced datasets. This study is one example of the potential integration of machine learning into predicting outcomes in Major League Baseball and provides a foundation for future studies. Additionally, this study was able to provide evidence towards the most

important statistics for predicting team success through the use of multiple feature selection algorithms. As somewhat of a surprise, there was no one particular algorithm that provided superior performance over the others, and it may be more beneficial to use them in conjunction as opposed to comparing the separate results of each. While the literature review and past work in this field were unable to provide any substantial support or confirmation of the results achieved in this study, they did provide some useful information and techniques to steer the direction of my research and improve the performance of the chosen machine learning models.

Next Steps

One of the major limitations of this study is that despite the data having a range of over fifty years, the sample size itself was fairly small. It is possible that the impressive performance results of the machine learning models for both the Winning Record and Playoffs classification problems are in part due to model overfitting. This is a risk of limited datasets, where the training data size is too small and does not contain enough data samples to accurately represent all possible input data values. A crucial next step will be to identify whether the findings in this study are legitimate not. Increasing the number of samples and rerunning the models could help clarify the results. This can be done by including additional seasons from Major League Baseball or including data from other baseball leagues that share the same statistics. The minor leagues and college baseball leagues in the United States, as well as baseball leagues in Mexico, Japan, and Korea could all provide additional data for evaluation. There is also the possibility that this was just simply an easy classification problem, and by conducting extensive testing on independent data, the consistency of the results can be validated. If these tests do return similar

results, perhaps a more complex multiclass problem could be formulated which could be used to classify team performance with even greater detail. Regression modeling can also be tested to not just predict a particular class of team success, but actual winning percentages.

Another limitation was the scarcity of related literature. While there are many published studies on the subject of machine learning in sports, most of it pertains to predicting individual game outcomes or player performance. These studies are also mainly focused on sports such as soccer and basketball. The inability to compare the results of this project with similar work means that while immediate evaluation and validation is difficult, it provides a large opportunity to expand on the work done so far. Future work can include conducting similar classification on different sports to confirm if comparable results are observed, or if the findings seen in this study are unique to baseball. Perhaps the application of the feature selection and machine learning methods used in this study are effective on a much larger dataset of individual baseball game results, as opposed to entire seasons. There is also the opportunity to combine the results of this study with the results of those seen in the literature review to predict future performances, instead of simply classifying past performance.

References

(2022). Baseball Reference. https://www.baseball-reference.com/

(2022). How Many Wins Will It Take To Make Expanded Playoffs? Baseball America. https://www.baseballamerica.com/stories/how-many-wins-will-it-take-to-make-expanded-playoffs/

(2020). MLB Team Payroll Tracker. Spotrac. https://www.spotrac.com/mlb/payroll/

(2020). Daily fantasy sports. Wikipedia. https://en.wikipedia.org/wiki/Daily_fantasy_sports#Growth

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. The Journal of Artificial Intelligence Research, 73, 1285-1322. https://doi.org/10.1613/jair.1.13509

Fullerton, Steven & Fullerton, Tom & Walke, Adam. (2014). An Econometric Analysis of the 2013 Major League Baseball Season. Research in Business & Economics Journal. 9. 115-120.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A Review of Data Mining Techniques for Result Prediction in Sports. Advances In Computer Science : An International Journal, 2(5), 7-12

Houser '05, Adam (2005) "Which Baseball Statistic Is the Most Important When Determining Team Success?," The Park Place Economist: Vol. 13 Available at: https://digitalcommons.iwu.edu/parkplace/vol13/iss1/12

Jain, P.K., Quamer, W. & Pamula, R. Sports result prediction using data mining techniques in comparison with base line model. OPSEARCH 58, 54–70 (2021). https://doi-org.ezproxy.lib.ryerson.ca/10.1007/s12597-020-00470-9

Nguyen, N., Ma, B., Hu, J. (2020). Predicting National Basketball Association Players Performance and Popularity: A Data Mining Approach. In: Nguyen, N.T., Hoang, B.H., Huynh, C.P., Hwang, D., Trawiński, B., Vossen, G. (eds) Computational Collective Intelligence. ICCCI 2020. Lecture Notes in Computer Science(), vol 12496. Springer, Cham. https://doi.org/10.1007/978-3-030-63007-2_23

Teno, G.D.S., Wang, C., Carlsson, N., Lambrix, P. (2022). Predicting Season Outcomes for the NBA. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds) Machine Learning and Data Mining for Sports Analytics. MLSA 2021. Communications in Computer and Information Science, vol 1571. Springer, Cham. https://doi-org.ezproxy.lib.ryerson.ca/10.1007/978-3-031-02044-5_11

Thabtah, F., Zhang, L. & Abdelhamid, N. NBA Game Result Prediction Using Feature Analysis and Machine Learning. Ann. Data. Sci. 6, 103–116 (2019). https://doi-org.ezproxy.lib.ryerson.ca/10.1007/s40745-018-00189-x