HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Evaluating Baseball Statistics by Predicting Playoff Teams

Rohan Nakra[1] and Ryan Kimes[1#]

[1]Huntington Beach High School
[#]Advisor

ABSTRACT

In this paper, we explore how different baseball statistics correlate to an entry to the playoffs. We use LogisticRegression and XGBoost to evaluate if a baseball statistic has a high correlation with whether or not a team makes the playoffs. We set up three models: the Moneyball model (which uses moneyball statistics), the All Stats Model (which uses moneyball statistics *and* additional common statistics), and the XGBoost model (which uses the same dataset of the All Stats model, but the structure of the model is different). We compare these models, evaluating their accuracy, variable coefficients, and confusion matrix. From these tests, we find that the Moneyball Model has similar accuracies to the All Stats Model, revealing that moneyball statistics are still a relevant and accurate way to predict if a team makes the playoffs. The variable coefficients test highlights that moneyball statistics have the highest importance in the model's ability to predict if a team makes the playoffs. While the tests provide a foundation for the evaluation of moneyball and common baseball statistics, there remains future opportunities to use different models and a larger dataset.

Introduction

There are many ways to evaluate success in baseball and thanks to the rise of artificial intelligence and machine learning, more and more tools are available for data analysts to solve problems in baseball analytics. As agreed upon in every sport, winning the championship is the goal for all teams; however, how to reach the World Series in baseball is a formula that teams are still trying to figure out. Although not a formula, former Chicago Cubs and Boston Red Sox President of Operations, Theo Epstein, simplifies the goal into three words: make the playoffs. In many ways Epstein is correct. Since 1995, only 3 who had the best regular season record and went on to win the World Series that same season. Epstein utilized analytics, resembling the ideas of the moneyball theory introduced by the Oakland Athletics in the 2002 season which is another concept we analyze in this paper. Oakland was a small-market team who always had difficulty making a run in the playoffs because they did not have the money to invest in big-market players. Instead, the Athletics used statistics (referenced as moneyball statistics in this paper) like on-base percentage to find players that had the production but were overlooked because they did not have the attributes of a star player. This new method initially was not well received but was eventually adopted by Major League Baseball (MLB) teams. More recently, studies have shown that moneyball statistics such as fielding independent pitching (FIP), on-base percentage (OBP), and slugging percentage (SLG) continue to accurately quantify defensive and offensive performance amid new statistics from the "Statcast" era (Mizels, Erickson, and Chalmers 2022). In this paper, we further explore the boundaries of moneyball statistics and other common baseball statistics to predict if a team makes the playoffs through different tests and machine learning models.

Methods

Data Creation and Preparation

Table 1. Stat Definitions

OBP	on-base percentage
SLG	slugging percentage
ERA	earned run average
ERA+	earned run average +
FIP	fielding independent pitching
HR	home runs
RBIs	runs batted in
RC	runs created
Ks	strikeouts
BA	batting average

There are many sources to retrieve baseball analytics data. This project uses a Kaggle dataset (link in supplemental materials) which displays all team stats from 1962 - 2012. These stats include OBP, SLG, and more; however, this dataset only focuses on moneyball statistics and excludes other common baseball stats (HR, RBI, Ks, etc.). Each sample in the dataset represents the stats of an MLB team from a specific season (Note that this dataset does not give statistics for an individual player during a specific year, it gives statistics for an entire team during a specific year). The target is a 1 or 0. 1 indicates the particular sample/team made the playoffs and 0 indicates the particular sample/team misses the playoffs.

The Kaggle dataset includes statistics such as wins and runs scored. These statistics would reveal too much to the model where it would recognize these statistics as a straightforward way to determine if a team made or missed the playoffs (if a certain team won enough games or scored enough runs, they would make the playoffs). However, comparing offensive and defensive statistics (listed in Table 1) can reveal patterns that differentiate playoff teams from non-playoff teams. For example, evaluating these statistics can answer questions: is batting average overrated? Do teams need to hit a high amount of home runs to make the playoffs? Do a high number of strikeouts always correlate to regular-season success? This is why we drop statistics such as wins and runs scored and instead focus on the offensive and defensive statistics in Table 1 to explore patterns in common baseball statistics.

Furthermore, the Kaggle dataset includes statistics such as opponent on-base percentage and opponent slugging percentage; however, both statistics contain undefined values for certain samples. As a result, we dropped both of these statistics from the dataset.

After the data manipulation in the previous two paragraphs, the dataset only includes three statistics: OBP, SLG, and BA; however, through the API sporsipy, we add additional stats to the dataset (ERA, ERA+, FIP, etc.). The Kaggle dataset includes data from the years 1962-2012. We also use the API to add data from the 2013-2021 MLB seasons.

Next, we split the dataset into two datasets. The "All Stats" dataset uses all the statistics listed in Table 1 while the "Moneyball" dataset uses only OBP, SLG, FIP, and RC (moneyball statistics).

Table 2. All Stats Dataset (first 3 samples)

Index	OBP	SLG	BA	ERA	ERA+	FIP	Ks	HR	RBIs	RC
0	.308	.373	.250	4.04	101	4.18	771	132	566	631.95
1	.335	.394	.271	3.55	121	3.81	914	137	707	737.18
2	.341	.441	.278	3.79	101	3.81	886	204	807	836.32

Table 3. Moneyball Dataset (first 3 samples)

Index	OBP	SLG	FIP	RC
0	.308	.373	4.18	631.95
1	.335	.394	3.81	737.18
2	.341	.441	3.81	836.32

Next, we use StandardScaler to scale the data values to minimize the range.

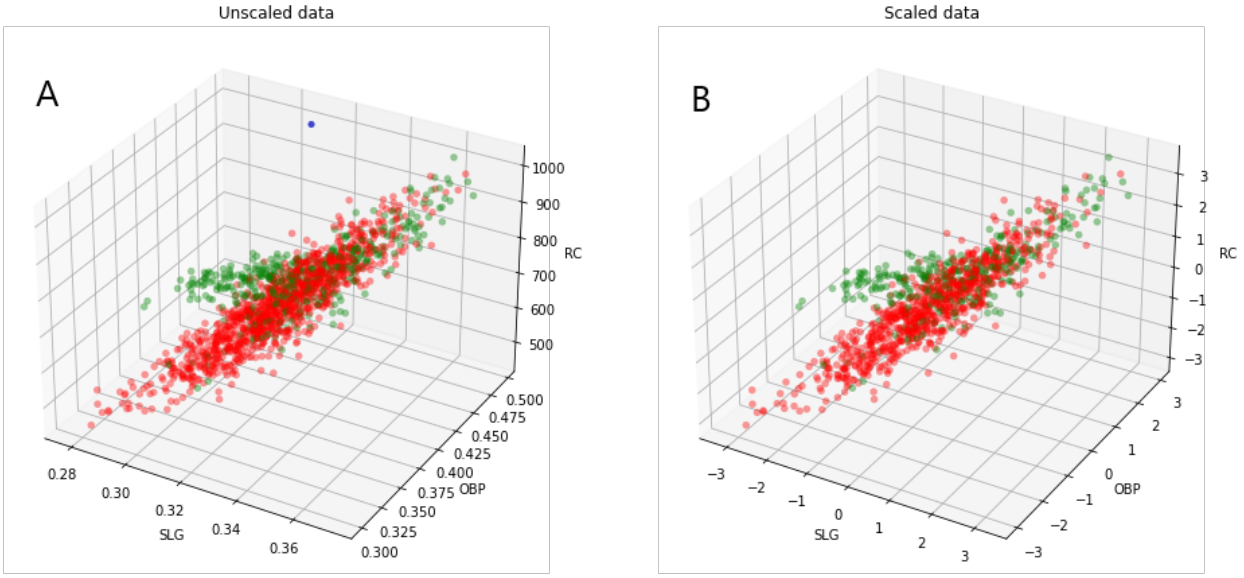

Figure 1. Unscaled vs Scaled Data. Red dots represent teams that missed the playoffs and green dots represent teams that made the playoffs (this will be the same legend used in succeeding figures). Notice that the values on the axes are now scaled in figure 1A. We use this scaler in order to eliminate potential bias from the models where they value one statistic over another. Without the scaled data, the models would have bias towards RC because the values are so high as shown in 1A.

Finally, we split the data into training and testing datasets. Think of the training dataset as a study guide and the testing dataset as an exam. The model “studies” the training dataset and then treats the testing dataset as an exam (where it cannot look at the answers/target… in this case, the target is whether or not a team makes the playoffs). Since the MLB has been around for 146 years and only a fraction of teams make the playoffs every year, there is a limited amount of data; however, we allocate 1104 samples to the training set and 368 samples to the test set. This balances the amount of data analyzed by the models.

Model Training

In this project, In this project, we use three different models that are trained on different subsections of the data. The first model is LogisticRegression which is trained on the "Moneyball" dataset (Moneyball model). The second model is LogisticsRegression but trained on the "All Stats" dataset (All Stats model). The final model is XGBoost trained on the "All Stats" dataset (XGboost model). The first two models are used to compare the effectiveness of moneyball statistics (if the Moneyball model and All Stats model perform similarly, it shows how moneyball statistics are still valuable in predicting playoff teams). The third model is used to compare the results of a linear model to a non-linear model (XGBoost is a non-linear model). None of the models use validation or other training techniques because this project compares basic models to reveal patterns within the data that we can interpret. Note that all models are from the sklearn python module and for the following tests, we use built-in features of the models along with metrics from the numpy and sklearn modules.

Results

Accuracy

The most basic test run on all models is an accuracy evaluation. The models will be referred to as the Moneyball model, All Stats model, and XGBoost model respectively.

Table 4. Model Accuracy

Model	Train Accuracy	Test Accuracy
Moneyball Model	77.26%	75.54%
All Stats Model	80.16%	79.89%
XGBoost Model	84.96%	81.52%

Note that both the Moneyball Model and All Stats Model have similar accuracies (even though the Moneyball Model only uses 4 statistics to determine if a team makes the playoffs). The XGBoost model also has a jump in accuracy compared to the two linear models.

Variable Coefficients

The LogisticRegression model utilizes coefficients to give a certain weight to variables which helps it classify a given sample. We can use these coefficients as a method to evaluate which variables (statistics) correlate most to a team making the playoffs. However, we only ran the variable coefficients test on the All Stats Model in order to compare

the coefficient values of the moneyball statistics to the coefficient values of *all* of the statistics used (the Moneyball model would return variable coefficients of strictly moneyball statistics and the XGBoost model is not linear and is unable to run the variable coefficients test). Note that values closest to 1 represent a high linear correlation to the target (making the playoffs).

Table 5. Variable Coefficients (All Stats model)

OBP	RBIs	SLG	ERA	BA	Ks	FIP	HR	ERA+	RC
1.025	.802	.673	1.368	.547	.536	.473	.402	.12	2.723

Notice the difference in values for OBP and BA. While BA is used more often to show a player's offensive production in today's game, OBP is a much higher indicator for this machine learning model to predict whether or not a team makes the playoffs.

Time Period Test

This test specifically trains and tests each model on a subset of the data from a specific time period. In this project, the models are trained and tested on data from 1962 - 1993 and then data from 1994 - 2021. We chose to partition the data this way because the early to mid-90s was when the steroid era started, a pivotal change in baseball history. While steroids were banned from the league in 1991, they had a pivotal impact on how the game was played, where the home run was valued more than ever.

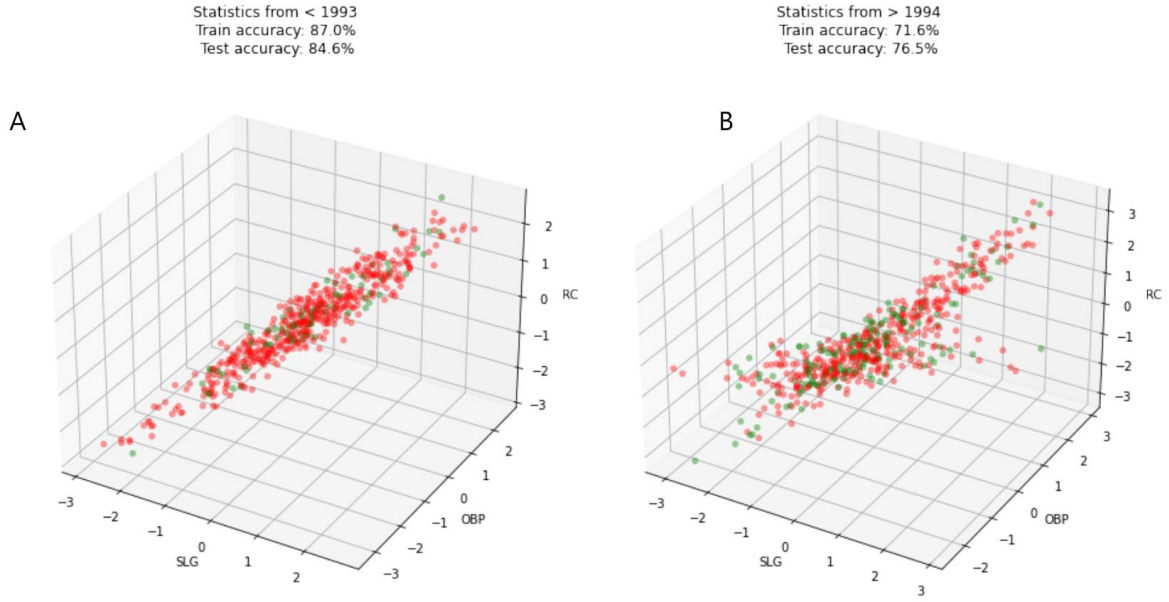

Figure 2. Time Period Test - Moneyball model. Notice the cluster of green and red dots in the second era, showing a more random pattern and a lower accuracy.

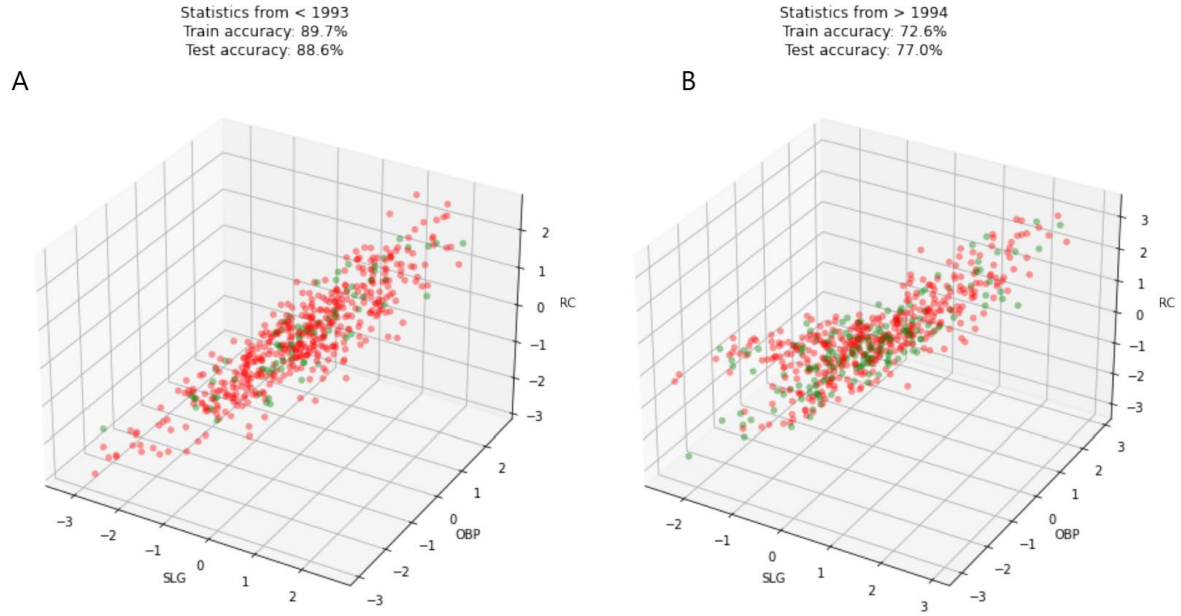

Figure 3. Time Period Test - All Stats model. Similar accuracies to Figure 2 reveal how the moneyball stats are the main stats used by LogisticRegressoin to classify samples.

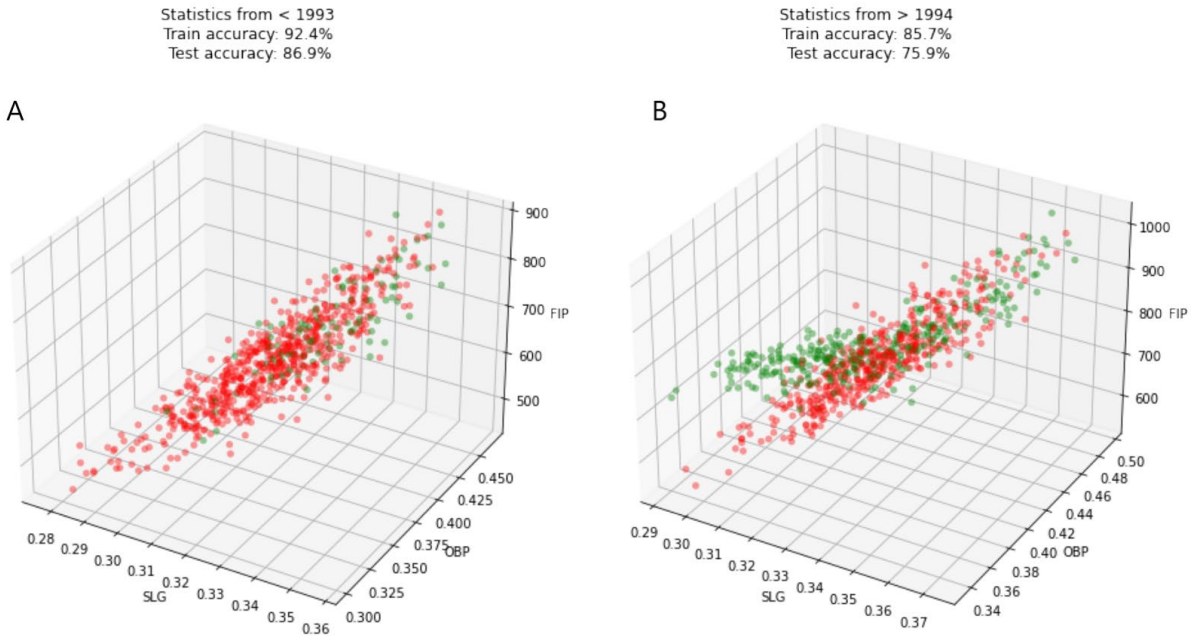

Figure 4. Time Period Test - XGBoost model. Slightly higher accuracies compared to Figures 2 and 3.

Furthermore, we also test the variable coefficients for the All Stats model time period test because this reveals coefficients for all variables/statistics. From this, we find that OBP and SLG have the highest linear correlation in *both* eras. RC has the third-highest linear correlation in the first era (before 1993), but the lowest linear correlation in the second era (after 1994). ERA and Ks are ranked at the bottom in the first era but jump into the top 4 ranked statistics in the second era.

Confusion Matrix

This test aims to evaluate the positive rate and negative rates of the models. Note that in this situation, the positive rate is the accuracy of a model to predict teams that made the playoffs correctly and the negative rate is the accuracy of a model to predict teams that missed the playoffs correctly. While both are important, the dataset contains significantly more teams that missed the playoffs than made the playoffs over the last 60 years of baseball. This is why we value the positive rate more than the negative rate.

As a guide, the way to read a confusion matrix is as follows:

Table 6. Confusion Matrix Blueprint

TN	FP
FN	TP

This represents a model for a confusion matrix, where TN is the number of negative samples predicted correctly, FN is the number of negative samples predicted incorrectly, TP is the number of positive samples predicted correctly, and FP is the number of positive samples predicted incorrectly. Note that positive is synonymous with a particular sample being labeled 1 (made the playoffs) and negative is synonymous with a particular sample being labeled 0 (missed the playoffs).

We can evaluate a model's performance in predicting a team that makes the playoffs (labeled 1) by dividing the number of samples that it predicted 1 correctly by the total number of samples predicted 1 (disregarding if the prediction was correct or not). Which would give the equation for the positive rate:

$$positive\ rate\ = \frac{TP}{TP + FP}$$

And in the same way, this would be the equation for the negative rate:

$$negative\ rate\ = \frac{TN}{TN + FN}$$

Table 7. Confusion Matrix - Moneyball Model

233	32
58	45

Table 8. Confusion Matrix - All Stats Model

240	25
49	54

Table 9. Confusion Matrix - XGBoost Model

245	20
48	55

From these matrices, we can now calculate and compare the positive and negative rate of each model.

Table 10. Positive and Negative Rates for each model

Model	Positive Rate	Negative Rate
Moneyball Model	58%	80%
All Stats Model	68%	83%
XGBoost Model	73%	84%

Similar to the previous two tests, there is a consistent rise in performance between the three models. In this case, we would value positive rate over negative rate because there are only a select few teams that make the playoffs every year. If a model can predict which team this is accurately, it shows a greater understanding of the difference between playoff and non-playoff teams. While the XGBoost and All Stats Models have a high positive rate, the Moneyball Model does not. These results somewhat contradict the conclusion derived from the accuracies in both the time period and accuracy tests.

Additional Test: Evaluating the Statistics by Predicting Playoff Outcomes

While the previous tests evaluated if the statistics could predict if a team makes the playoffs, this test expands further and evaluates if the statistics can predict if a team wins the World Series. We take two of the highest correlated statistics from the test in section 3.2, OBP and ERA (we use ERA in this instance to use one offensive and one pitching statistic). The graph plots regular season OBP vs ERA for teams that made the playoffs (2000-2019). The green dots are the World Series champions for that season. On the right, there is another graph that plots using the same criteria but only includes the 2022 playoff teams.

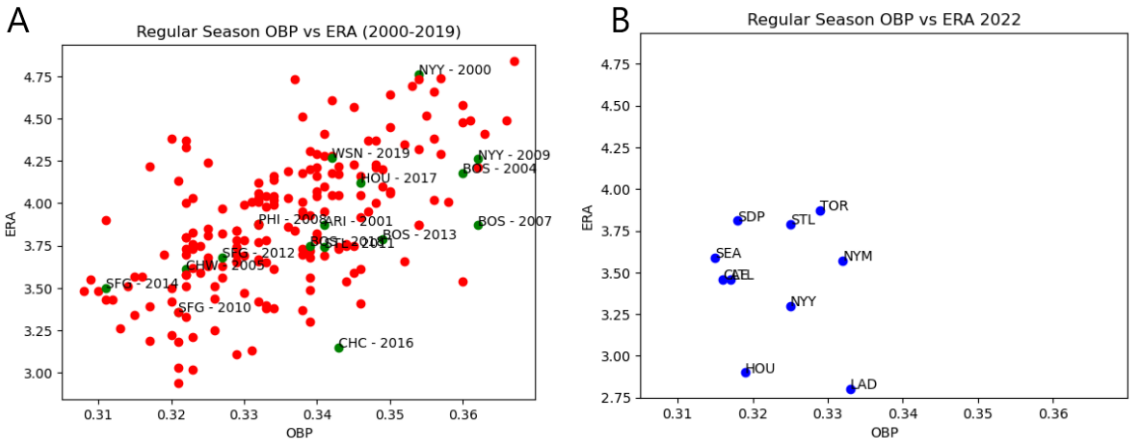

Figure 5. Comparing OBP and ERA trends for playoff teams. No clear pattern for playoff teams from 2000 - 2019 (teams that win the World Series do not always have a higher OBP and lower ERA. Notice how the scale of the axes in the second graph is the same as the first… the Dodgers look like heavy favorites. Both axes are labeled the same for direct comparison of Figures 5A and 5B

Additional Test: Evaluating Offensive Statistics by Predicting Playoff Outcomes

This test contains the same criteria as the previous test but uses offensive stats.

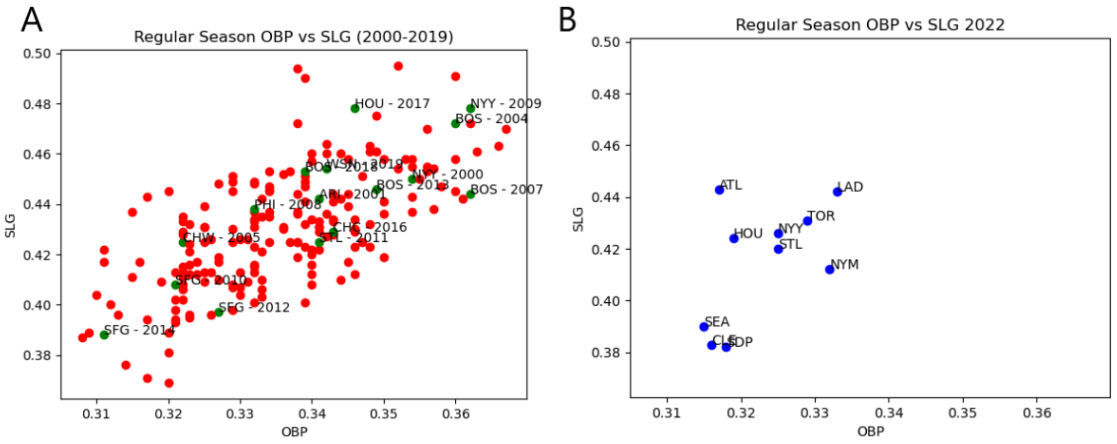

Figure 6. Comparing OBP and SLG trends for playoff teams. Again, no clear patterns for playoff teams from 2000 - 2019 exist. Both axes are labeled the same for direct comparison of Figures 6A and 6B.

Additional Test: Evaluating Pitching Statistics by Predicting Playoff Outcomes

This test contains the same criteria as the previous test but uses pitching stats.

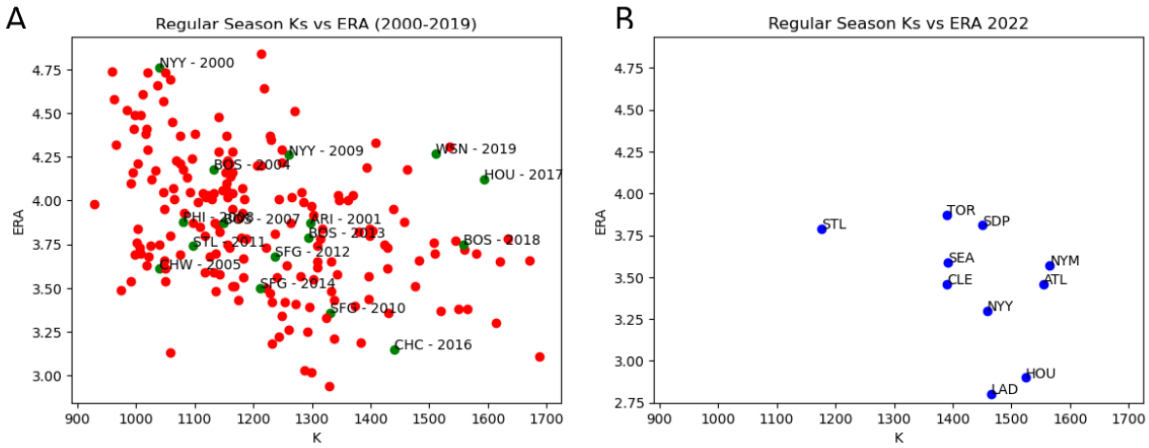

Figure 7. Comparing Ks and ERA trends for playoff teams. Again, no clear patterns for playoff teams from 2000 - 2019 exist. Both axes are labeled the same for direct comparison of Figures 7A and 7B.

Discussion

All models did well in all performance metrics. Accuracy shows a general increase within the three models which is possibly a result of the complexity of the models; however, the difference in accuracy is minimal, showing that even moneyball statistics alone prove to be an accurate way to predict a ticket to the playoffs.

The variable coefficients test also highlights the consistency of OBP and SLG to predict successful regular season teams. Both of these tests show the moneyball theory is still prevalent in today's game even when evaluated against other common baseball statistics. Compared to the second-ranked statistic in the variable coefficients test, RBIs, OBP has a linear correlation 8 times greater than the target. The fact that OBP ranks significantly higher than RBIs may come as a surprise considering that RBIs could be labeled as a statistic that "reveals" too much to the model; similar to wins or runs (see section 2.1 paragraph 2). This is because RBIs are runs battened in (not those that are a result of an error). Ks and HR rank lower which is ironic considering that baseball players are evaluated by most fans by the number of home runs they hit or the number of batters they strike out. RC, a statistic created by the creator of moneyball theory (Bill James), ranks the lowest which may be because it was initially created to analyze individual performance, not team performance.

The time period test does a better job of differentiating the models and the statistics. While the linear models have similar accuracies in both eras, the XGBoost model obtains a higher accuracy possibly because of its non-linear computation method and model complexity. Note that all models have significantly higher accuracies in the first era (1962 - 1993) compared to the second era (1994-2021). This can relate to a number of changes in baseball that occurred after the 1990s: the introduction of 8 playoff teams instead of 2 (this would later be increased to 10 in 2012), an increase in the number of home runs hit, and an increase in the number of strikeouts. These changes all contribute to baseball becoming more unpredictable, where one swing of the bat can change a crucial game where playoff hopes are on the line. The change in playoff structure can also be visualized by comparing Figure 2A and Figure 2B, for example. The former shows a small number of green dots (teams that made the playoffs) while the latter shows a much

larger number of green dots, accounting for the introduction of more playoff teams in the 1990s. We then run a variable coefficients test for the All Stats model to compare the results from the first and second eras. The jump from ERA and Ks (as noted at the end of section 3.3) made in the second era compared to the first could point to the higher strikeout rate in today's MLB and the rising number of "ace" pitchers who throw the ball faster than ever before.

The confusion matrix test underlines the additional statistics that increase the positive rate in the All Stats model and the XGBoost model. The Moneyball model struggles in classifying teams that make the playoffs correctly and based on the results found in Table 5 for the variable coefficients test, one could deduce that lacking crucial statistics such as RBIs and ERA hinders this model's ability to predict if a team made the playoffs. This test does a great job of revealing a model's true understanding of the dataset.

From the additional tests, we can see that randomness in the playoffs is certainly true. There is no clear pattern or trend for any of the additional tests. All a team needs is to get in the postseason... after which anything can happen. And even at the time of writing, the historic 2022 Los Angeles Dodgers with too many All-Stars to name have just lost to the San Diego Padres. This marks the largest postseason upset since 1906 with a 22-win difference between the two teams in the regular season.

We also run another additional test where we include the year the team played for in the dataset for each model. This means the Moneyball model would analyze a dataset containing the statistics OBP, SLG, FIP, RC, and Year (the same methodology for the All Stats model and XGboost model). Our hypothesis before running the tests was that accuracies would significantly increase considering that giving the models the year the team played for would give a greater context to the meaning of the statistics. This is because baseball has changed drastically between 1962 and 2021 as shown by the increase in strikeouts and home runs. For example, if the model was given a sample of a team that played in 1962 which had a low number of strikeouts and home runs, if it knows what year the team played in, it could better understand the context of *why* these numbers are so low and use it as a benchmark to compare against other teams in the same year/era. After running all three tests (accuracy, variable coefficients, and time period test), there were only minimal increases in performance across all tests (< 1%); however, Year was the third-ranked statistic in the variable coefficients test. The underwhelming results could be because the model is able to "contextualize" the statistics itself through StandardScaler and does not necessarily require Year.

Conclusion

This study further reaffirms that moneyball statistics remain relevant because of their ability to accurately indicate success in terms of a team making the playoffs (Mizels, Erickson, and Chalmers 2022). While baseball will continue a transition with "Statcast" statistics such as excepted ERA (xERA), exit velocity, and more, the effect moneyball statistics have had on the game should always be remembered.

Future work should incorporate a more comprehensive evaluation of these statistics through different machine learning models, more tests, and a larger dataset. While this research uses a LogisticRegression model to get the variable coefficients other models can be used to find different patterns in the statistics. For example, a polynomial regression can reveal if certain statistics follow a non-linear pattern. A larger dataset should include more statistics from years before 1962. This would allow for a time period test that could compare 3 or 4 different eras, not just 2. A new dataset could also include more advanced statistics such as On-base Plus Slugging Plus (OPS+), Batting Average on Balls in Play (BABIP), Defensive Runs Saved (DRS), and more. However, utilizing these advanced statistics must be used carefully as some are relatively new and inapplicable to older eras of baseball. Future work can also evaluate how risk correlates to regular season success. For example, if stolen bases and stolen base attempts correlate to a team making the playoffs.

It is hoped that this research will help better evaluate baseball statistics, understand how baseball has changed, and which statistics correlate to this change.

Limitations

Finding enough data was the main limitation of this project. As stated on page 3, There were only 1472 total samples that were used while normal machine learning models use hundreds of thousands of samples. A possible way to avoid the lack of data issue is to manipulate the dataset to give statistics for individual players which would significantly increase the number of samples available to use; however, evaluating player data would shift the objective of such a project toward player evaluation instead of evaluating the efficiency of baseball statistics.

Acknowledgements

I would like to thank my parents, Neal and Michelle Nakra, and Professor Sung-Jae Lee for helping me throughout this journey. I would also like to thank Summer Institute for the Gifted at UCLA, where I started this project.

References

Apostolou, K., & Tjortjis, C. (2019). Sports analytics algorithms for performance prediction. 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). https://doi.org/10.1109/iisa.2019.8900754

Appelman, D. (2008, March 15). Get to know: Runs created. FanGraphs Baseball. https://blogs.fangraphs.com/get-to-know-runs-created/

Cabral, C. (2020, June 23). Bill James: How sabermetrics changed baseball. Shortform Books. https://www.shortform.com/blog/bill-james-moneyball/

Chicago Cubs playoff history: 1885 - 2023. Playoff History | 1885 - 2023. (n.d.). https://champsorchumps.us/team/mlb/chicago-cubs

Longest MLB postseason droughts, active and all-time. , Active and All-Time. (n.d.). https://champsorchumps.us/drought/longest-mlb-playoff-drought#tab-drought-historic

Mizels, J., Erickson, B., & Chalmers, P. (2022). Current state of data and analytics research in baseball. Current Reviews in Musculoskeletal Medicine, 15(4), 283–290. https://doi.org/10.1007/s12178-022-09763-6

Taylor, B. (2014, June 8). Theo Epstein wants playoffs or suck, nothing in between, and other bullets. Bleacher Nation | Chicago Sports News, Rumors, and Obsession. https://www.bleachernation.com/cubs/2013/02/27/theo-epstein-wants-playoffs-or-suck-nothing-in-between-and-other-bullets/

Wade, C. (2020). Getting Started with XGBoost in scikit-learn. Medium. https://towardsdatascience.com/getting-started-with-xgboost-in-scikit-learn- f69f5f470a97