A Statistical Analysis of the MLB Strikeout Rate

Executive Summary

Samuel Northam

Master of Science, Data Analytics, Western Governors University

Dr. Daniel Smith, PhD

October 2022

Problem Statement

The MLB has been around since the late 1800s but the first World Series was not played until 1903. Throughout baseball's long history, statistics have always been at the forefront of the game. Teams have used statistics to set lineups and upper management uses advanced analytics to make organizational decisions.

This study focuses on the MLB strikeout rate, which is the total number of strikeouts per 100 at bats: $\frac{\mathrm{K}}{\mathrm{AB}} * 100$. The study analyzes historical records to answer the question "What extent can the Strikeout Rate in the MLB can be forecasted using Time Series Analysis?" The hypothesis of the study is "The Strikeout Rate in an MLB Season *can* be forecasted with 90% accuracy".

Data Analysis Process

Data was collected from baseball-reference.com using the *pybaseball* python library. Season totals were collected for all players for every season from 1903 to 2022, resulting in a csv file with over 87,000 entries. Single season statistics were then computed using the *Pandas* library.

The seasons from 1903 to 1912 were missing lots of entries in the *Strikeout* column, so data for these seasons was downloaded manually from baseball-reference.com. The *groupby* function form the *Pandas* library was used to get season totals. After acquiring the total at bats and strikeouts for every season from 1903 to 2022, the MLB Strikeout Rate was calculated. The

MLB Strikeout Rate is the total number of strikeouts per 100 at bats: $\frac{\mathrm{K}}{\mathrm{AB}} * 100$. The resulting dataframe had 120 rows and 2 columns: *season* and *rate.*

The data was then split into two sets: a training and a testing. The training dataset contained the data from 1903 to 1986 and the testing dataset contained data from 1987 to 2022. Time Series Analysis was then performed in order to find the best ARIMA model for the dataset. ACF and PACF plots suggested an MA 2 or 3 and AR 2 or 3 component for the series. There were not enough datapoints to conduct a seasonal composition and since the frequency of the dataset is yearly, no seasonal components were considered for the model. The *auto_arima* function was used to test and compare hundreds of models. This function used AIC corrected to compare models.

Findings

The result of the analysis was that an ARIMA(2, 1, 3) was the optimal model to forecast the MLB Strikeout Rate. A forecast for the 1987 to 2022 seasons were compared against the testing dataset to see how accurate the model was. Figure 1, on the following page, shows the forecast for these years compared to the observed. Forecasts were calculated using the *get_forecast* function. The MAPE of the model was 13.65%. A forecast for the 2023 to 2032 seasons were then calculated. Figure 2 shows the 10-year forecast.

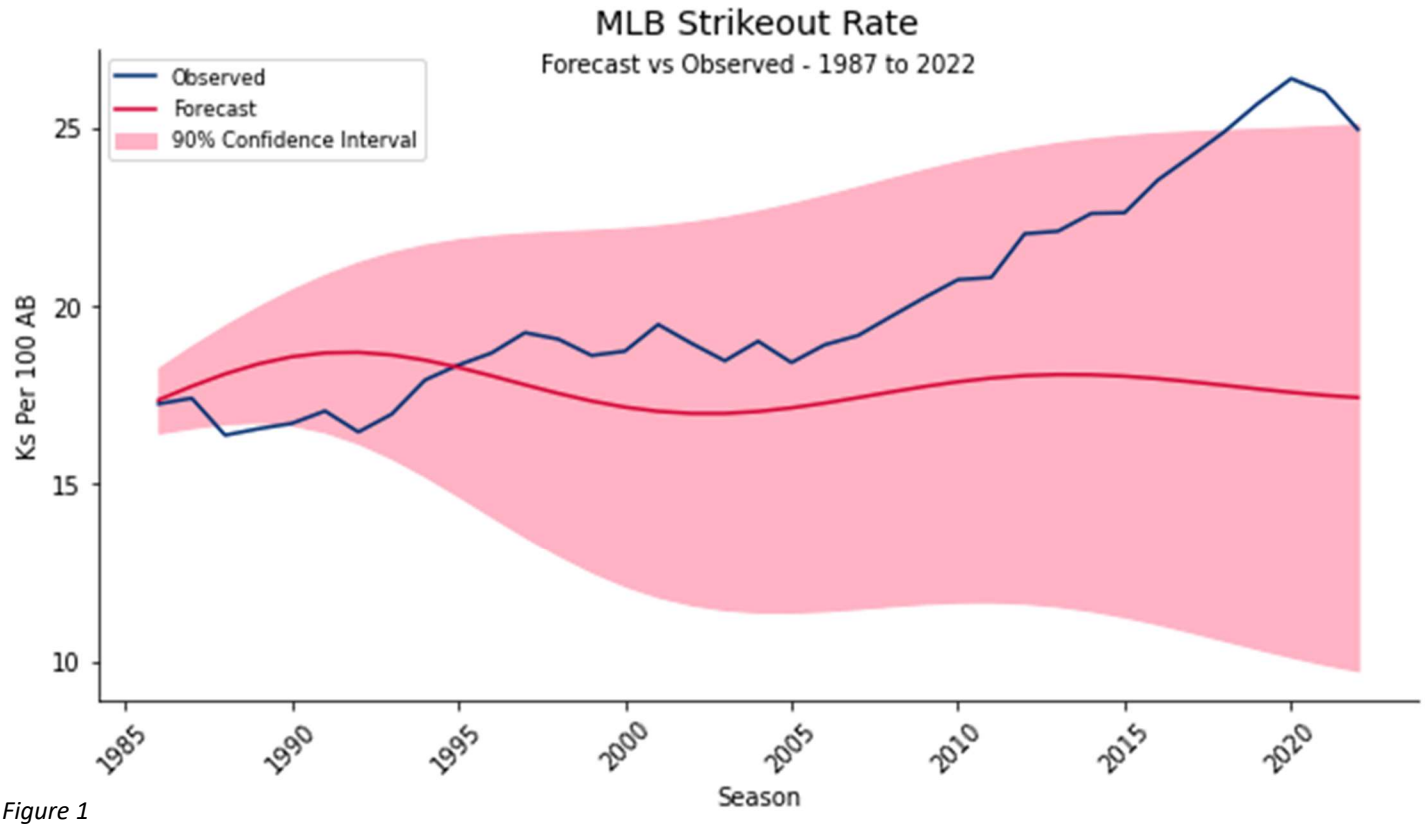

Figure 1

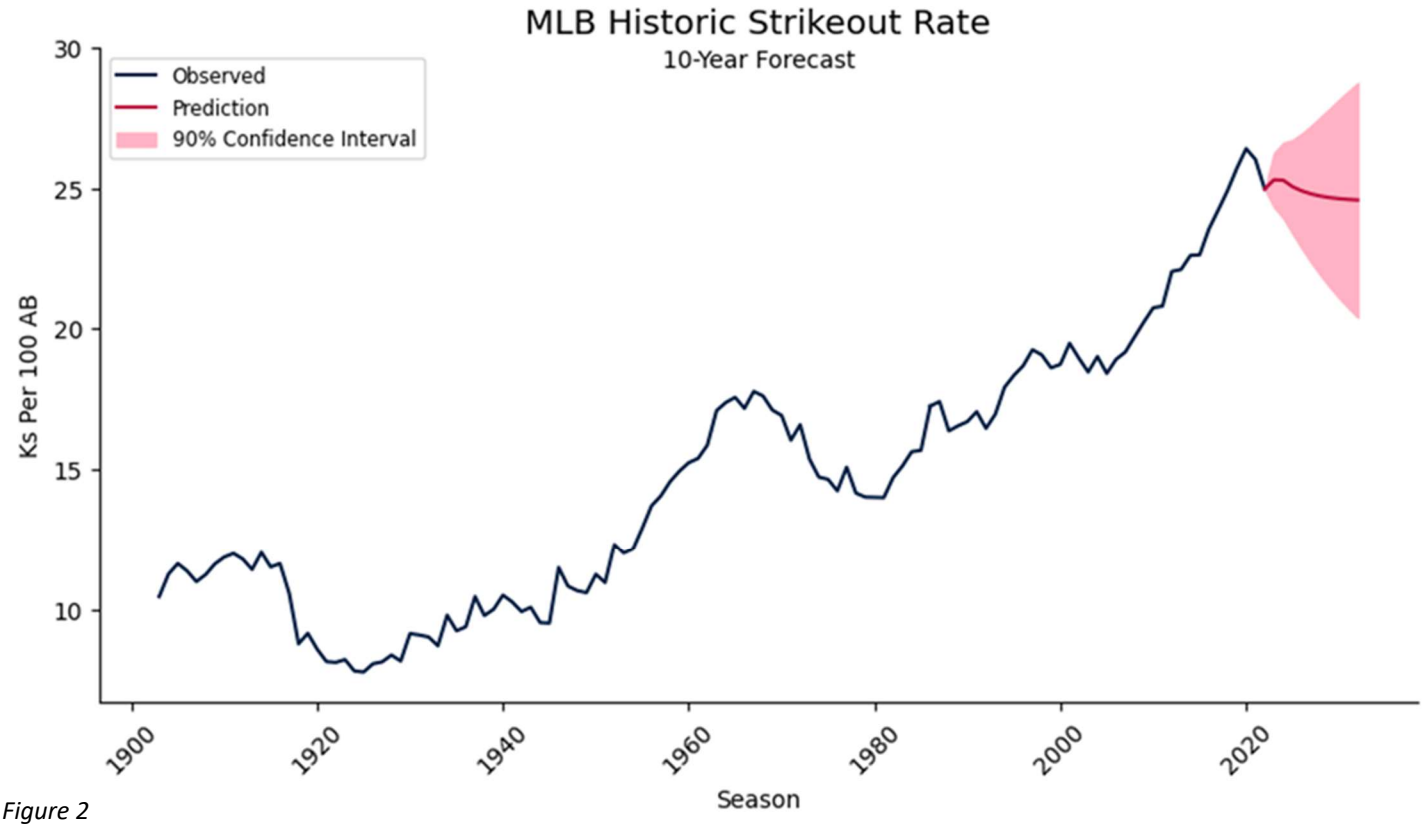

Figure 2

The mean prediction for the 10-year forecast stays fairly consistent at around 25 Strikeouts per 100 At Bats. This is due to the model failing to capture the true variability of the MLB Strikeout Rate which creates a fairly flat forecast. This prediction seems logical as the MLB Strikeout Rate was recently increasing at a rapid rate and reached a historic high before the MLB introduced multiple rule changes regarding pitchers. MLB hoped that this rule change would reduce strikeouts and increase total offense. With these rule changes, the strikeout rate will likely reduce back to normal levels before leveling out again.

Limitations

The main limitation in this analysis is the number of datapoints. The lack of datapoints led to a model that did not accurately capture the season-to-season variance of the MLB Strikeout Rate. Another limit is that the data was taken at a yearly frequency. Time Series Analysis would likely perform much better with biannually, quarterly, or weekly frequencies when dealing with a time period of this size. Including more datapoints would result in a model that more accurately captures the trend and variance of the MLB Strikeout Rate.

Proposed Actions

One suggestion for further research would be to recreate this study with weekly measurements rather than yearly. Weekly data would show how the MLB Strikeout Rate varies throughout the season and would give far more datapoints for the model to train on. It would

also allow for Seasonal Components to be included in a SARIMA model which would result in a model that could more accurately and precisely forecast the MLB Strikeout Rate.

Another suggestion would be to use biannual measures of the strikeout rate. A model may be able to use MLB Strikeout Rate at the All-Star game to predict what it will be once the season ends. This would also give the model double the number of datapoints, leading to a more accurate model.

Expected Benefits

The MLB could compare the Strikeout Rate the next 10 years to the forecast that the model made to see the effect of rule changes. Since the MLB wants to see an increase in offense, they will likely make more and more rule changes to benefit the offense. If despite these changes the Strikeout Rate continues to increase relative to the models prediction, the MLB will know that their changes were ineffective. This will allow the MLB to know if they need to switch strategies before it is too late.