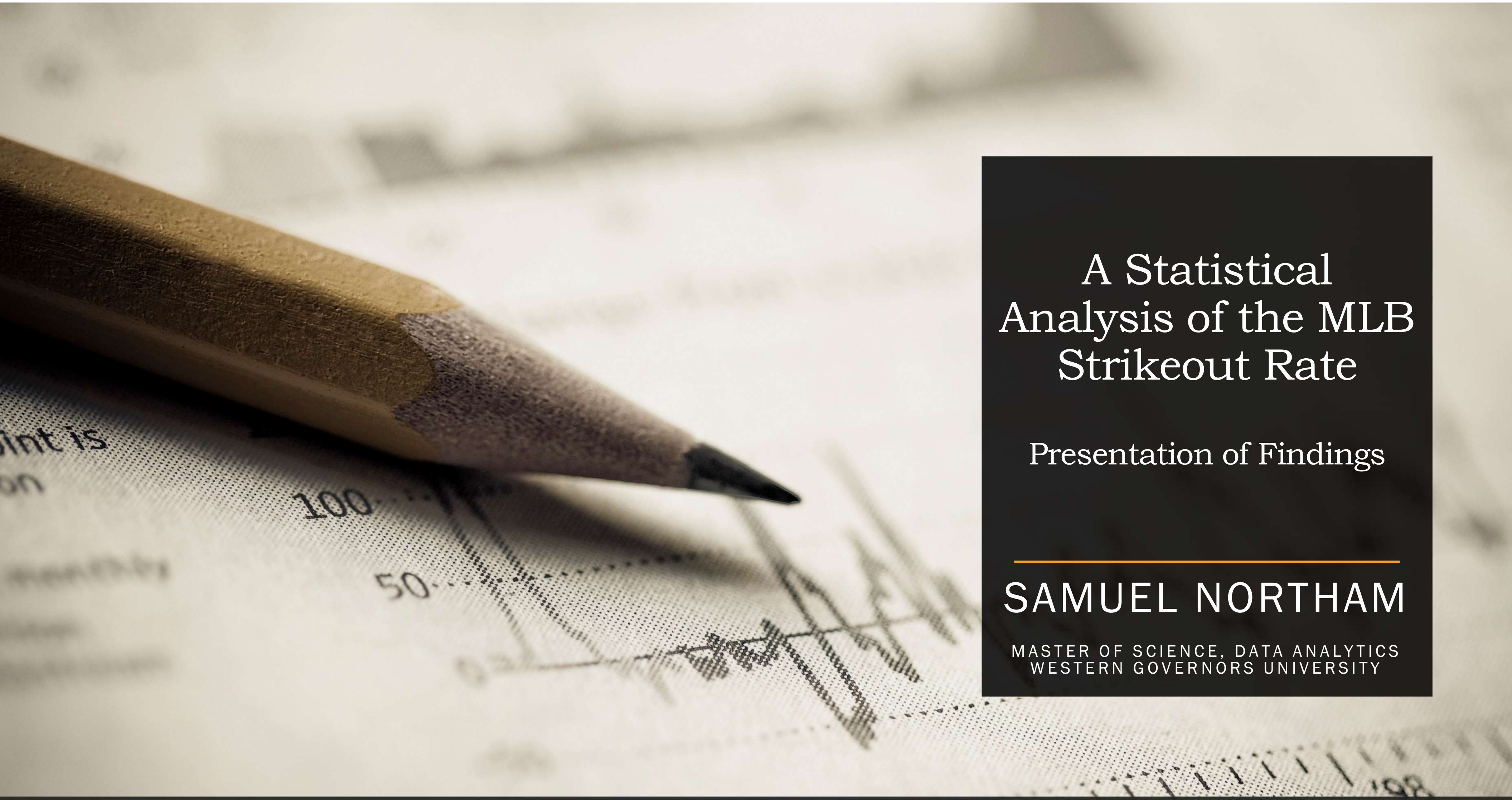


A Statistical Analysis of the MLB Strikeout Rate

Presentation of Findings

SAMUEL NORTHAM

MASTER OF SCIENCE, DATA ANALYTICS
WESTERN GOVERNORS UNIVERSITY

Introduction

- Graduated from Northern Arizona University in 2022
  - Studied Statistics and Computer Science

- Enrolled in a Data Analytics Masters Program at Western Governors University

Problem

- Statistics have been the lifeblood of baseball since its inception in the late 1800s
- Players use stats to give them an edge over their opponent, managers use them to set lineups, and The Major League Baseball uses them to help dictate rule changes
- The MLB Strikeout Rate has been steadily on the rise since the first World Series in 1903
- The MLB would benefit from a time series analysis of the strikeout rate to help them understand the trends and help them determine the success of rule changes

Research Question

- What extent can the Strikeout Rate in the MLB be forecasted using Time Series Analysis?

- Hypothesis: The Strikeout Rate in an MLB Season can be forecasted with 90% Accuracy

Data Acquisition

- Data was scraped from baseball-reference.com using the *pybaseball* python library
  - Scraped season level strikeouts (K) and at bats (AB) for every player from 1903 – 2022
  - Dataset contained over 87,000 rows

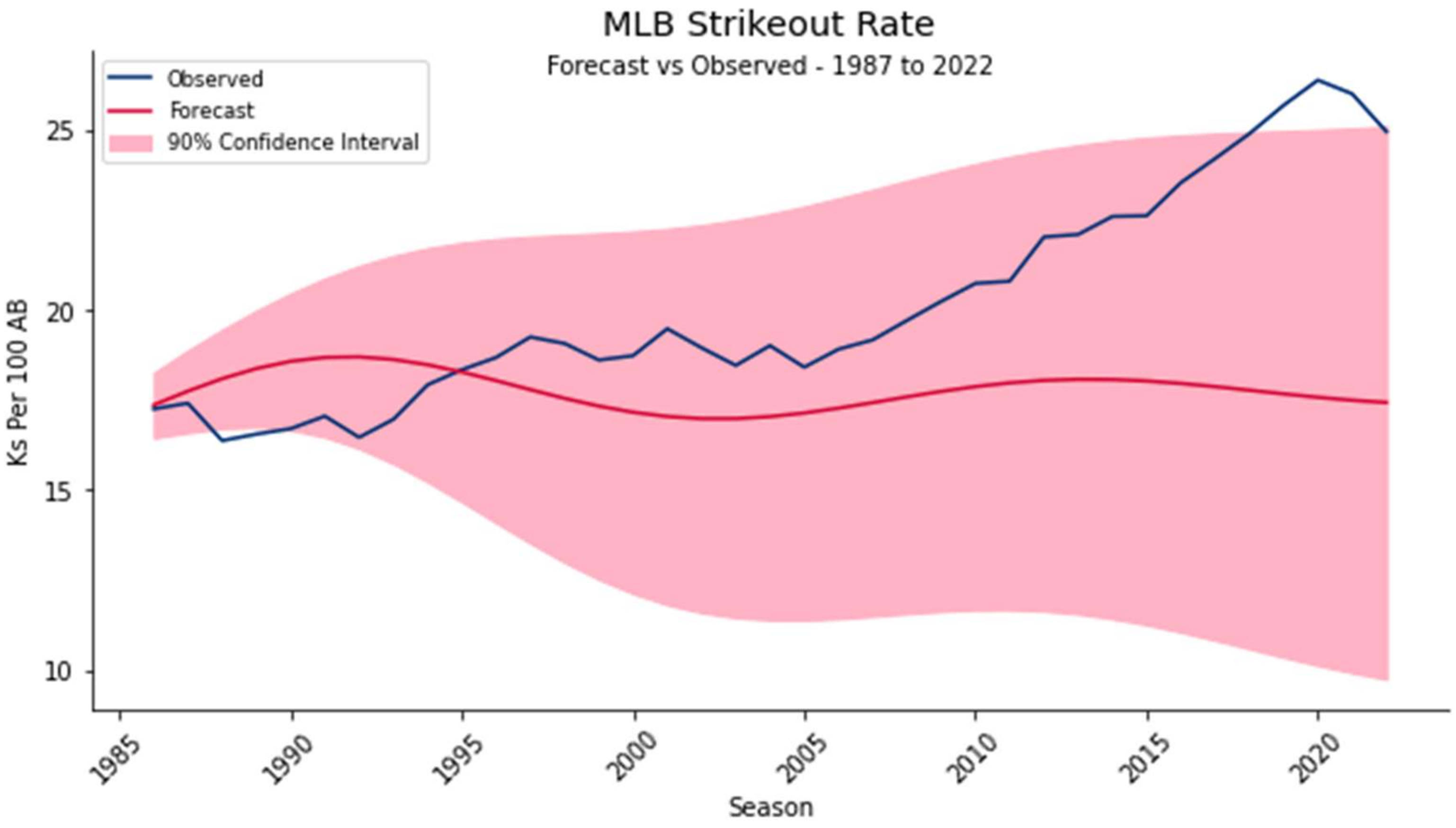
- League totals were calculated for every season using the *Pandas* python library
  - Calculated strikeout rate for every season
    - $\frac{\mathrm{K}}{\mathrm{AB}} * 100$

Data Analysis

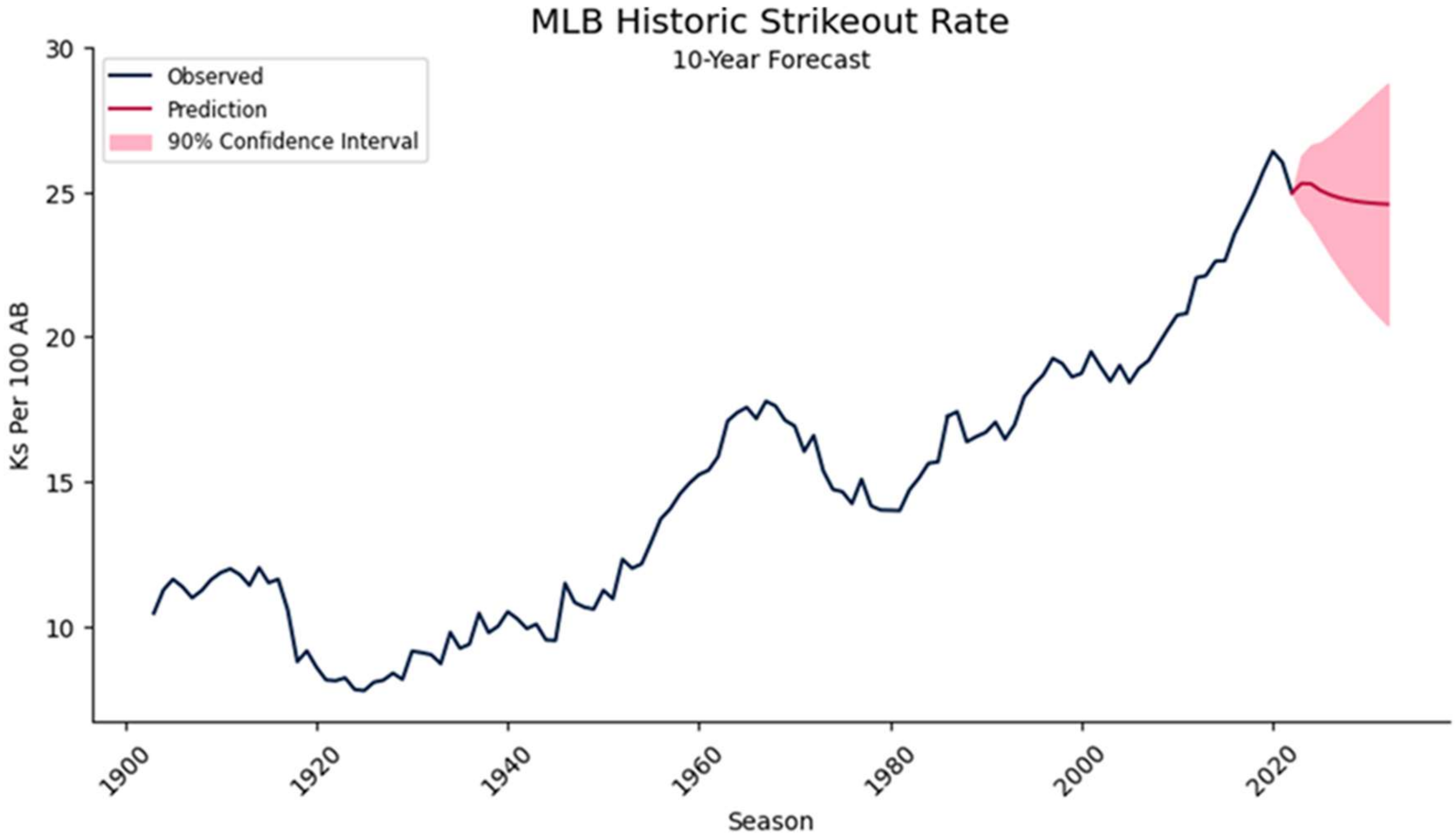
- Time Series Analysis was performed to find the optimal Auto Regressive Integrated Moving Average (ARIMA) model for the dataset
- ARIMA model was trained on the 1903 to 1986 seasons
- A 36-year forecast was calculated for the 1987 to 2022 seasons
  - Forecasted values were compared with observed values to score the model
- A 10-year forecast was generated for the 2023 to 2032 seasons

Findings

- Mean Absolute Percent Error (MAPE) was used to score the model
- MAPE of the final model suggested sufficient accuracy
- Confidence Interval of 36-year forecast nearly encapsulated all observed values
- 10-Year forecast predicts Strikeout Rate to slowly decrease and even out



MLB Strikeout Rate
Forecast vs Observed - 1987 to 2022
Observed
Forecast
90% Confidence Interval
Ks Per 100 AB
25
20
15
10
1985
1990
1995
2000
2005
2010
2015
2020
Season


MLB Historic Strikeout Rate
10-Year Forecast
Observed
Prediction
90% Confidence Interval
Ks Per 100 AB
30
25
20
15
10
1900
1920
1940
1960
1980
2000
2020
Season

Limitations

- Lack of Data
  - Time Series Analysis requires a large volume of data to properly capture trend
  - Full dataset only contained 120 datapoints
- Frequency of Data
  - Data was collected at yearly intervals
  - Made seasonal trend difficult to capture

Proposed Actions

- Obtain data at weekly intervals
  - Model will capture the season-to-season trend better with more data points

- Use biannual measurements
  - Can build a predictive model that uses the Strikeout Rate at the all-star break to predict the end of season strikeout rate

Benefits

- The findings of this study have the potential to help the MLB with future decisions

- MLB could compare the Strikeout Rate the next 10 years to the forecast to see the effect of rule changes
  - MLB wants to see an increase in offense

- MLB has implemented rule changes to increase offense
  - Can compare actual values to the forecast to see if the effects of the rule changes are noticeable
  - Can see if new rules need to be implemented

References

◦ Chatfield, C. & Xing, H. (2019). The analysis of Time Series: An introduction with R. Chapman & Hall/CRC.

◦ Jldbc. (2022). JLDBC/Pybaseball: Pull current and historical baseball statistics using Python (Statcast, Baseball Reference, fangraphs). GitHub. Retrieved October 17, 2022, from https://github.com/jldbc/pybaseball