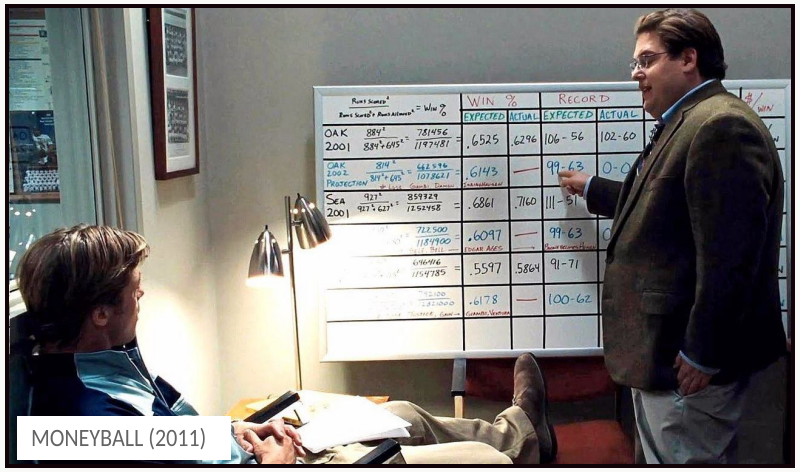
STEPPING UP TO THE PLATE

Predicting MLB Player Value & Team Wins with Machine Learning

Eric Au

BUSINESS UNDERSTANDING

THE FIRST PITCH



MONEYBALL (2011)

SOURCE: FORBES

- **How does a team determine the monetary value of a player?**
  - Teams are structured around how they will pay their most valuable players.
- **Player's Salary = Player's Value?**
  - Value is subjective
  - Predict player salary based on performance
- **Winning In a Game of Failure**
  - Predicting team wins based on team metrics

BY THE NUMBERS

"Baseball has always been filled with negative statistics" - Joe Torre

DATA SOURCES:

- FanGraphs, Baseball Reference

STATCAST POWERED BY Google Cloud

BATTERS

- **Basic**: 9100 x 28 Features (2000 - 2021)
- **Advanced:** 3200 x 321 Features (2014 - 2021)

PITCHERS

- **Basic:** 9400 x 34 Features (2000 - 2021)
- **Advanced:** 3900 x 334 Features (2014 - 2021)

TEAMS

- **Basic:** 1600 x 61 Features (1960 - 2021)
- **Advanced:** 210 x 634 Features (2014 - 2021)

Name

Mike Trout (LAA, OF) $37,116,667

Nolan Arenado (STL, 3B) $35,000,000

Manny Machado (SDP, 3B) $32,000,000

Miguel Cabrera (DET, 1B) $30,000,000

Jose Altuve (HOU, 2B) $29,000,000

Giancarlo Stanton (NYY, OF) $29,000,000

Albert Pujols (LAD, 1B) $29,000,000

Anthony Rendon (LAA, 3B) $28,071,428

Bryce Harper (PHI, 1B) $27,538,462

Paul Goldschmidt (STL, 1B) $26,000,000

Joey Votto (CIN, 1B) $25,000,000

George Springer (TOR, OF) $23,666,667

2021's HIGHEST PAID POSITION PLAYERS

Source: FanGraphs

Pitcher Name

Trevor Bauer (LAD, P) $40,000,000

Gerrit Cole (NYY, P) $36,000,000

Max Scherzer (LAD, P) $35,920,616

Jacob deGrom (NYM, P) $35,500,000

Zack Greinke (HOU, P) $35,000,000

Stephen Strasburg (WSN, P) $35,000,000

David Price (LAD, P) $32,000,000

Clayton Kershaw (LAD, P) $31,000,000

Chris Sale (BOS, P) $30,000,000

Patrick Corbin (WSN, P) $24,416,667

Zack Wheeler (PHI, P) $22,500,000

Yu Darvish (SDP, P) $22,000,000

Johnny Cueto (SFG, P) $21,833,334

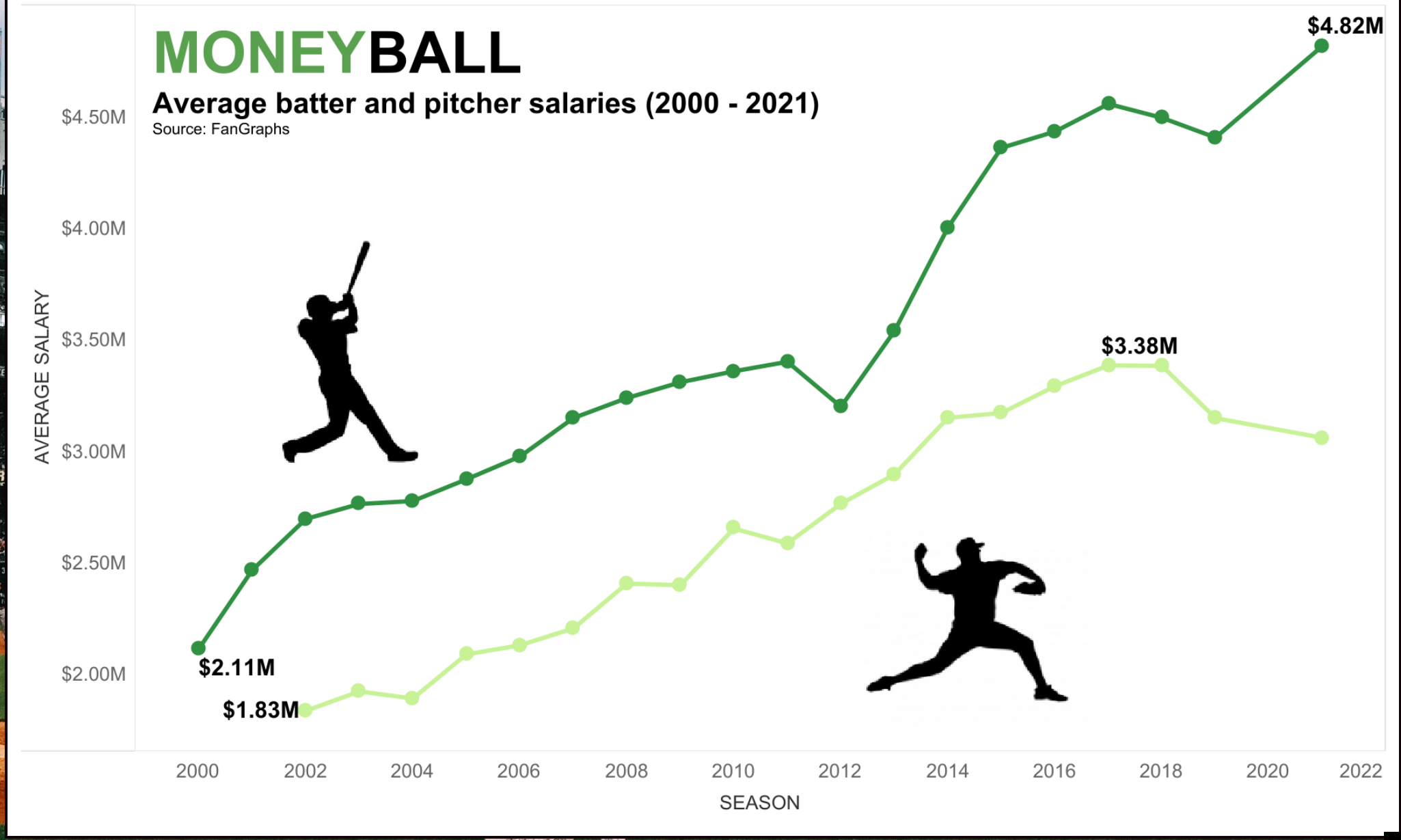
2021's HIGHEST PAID PITCHERS

Source: FanGraphs

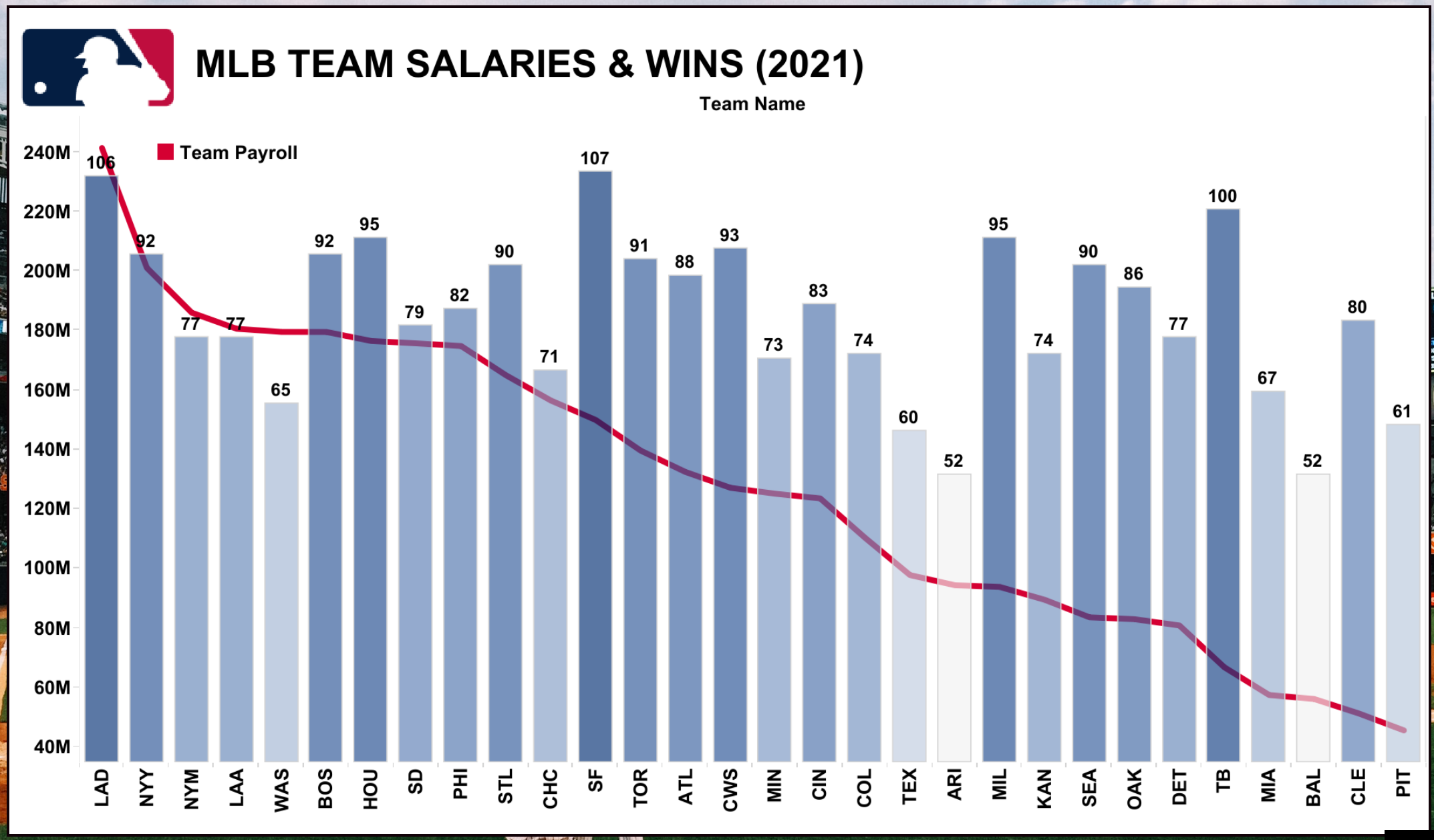
MONEYBALL

Average batter and pitcher salaries (2000 - 2021)

Source: FanGraphs

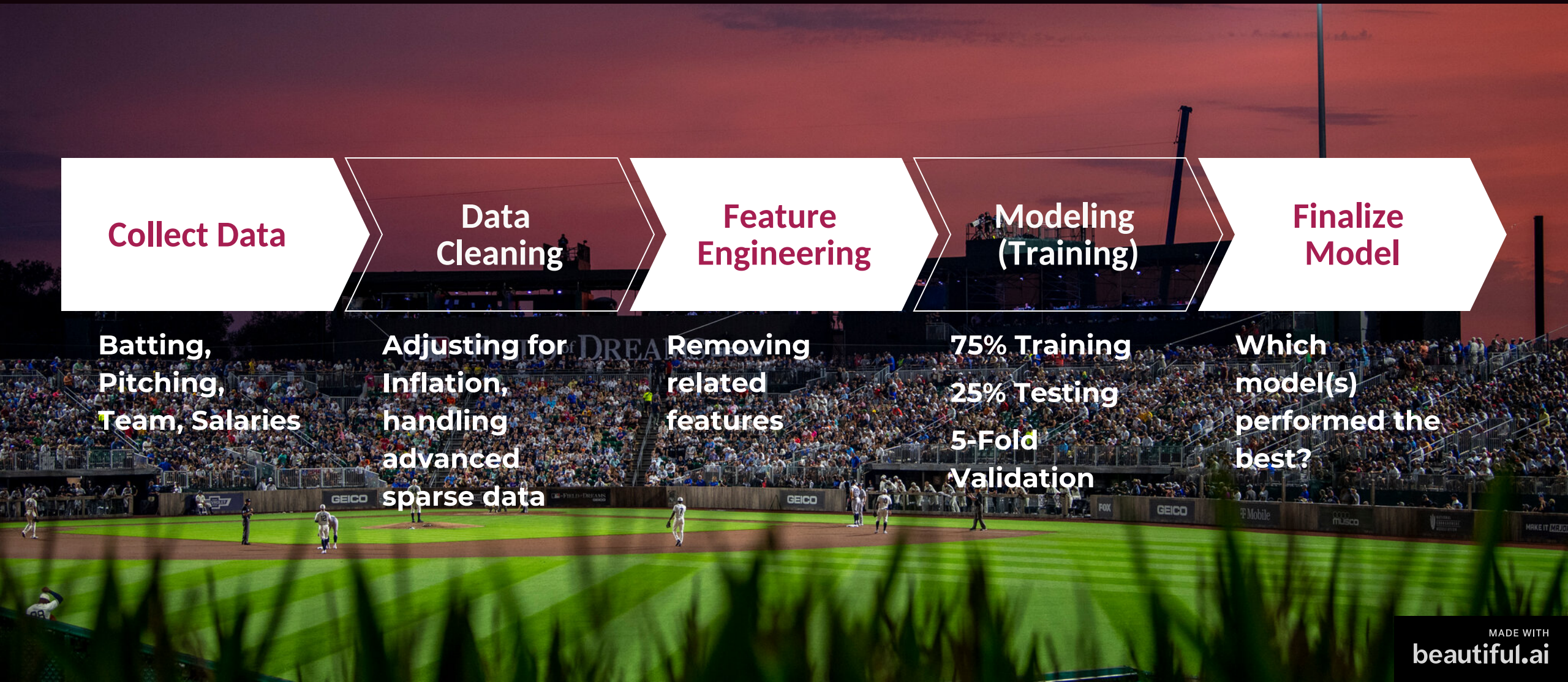


MLB TEAM SALARIES & WINS (2021)



IF YOU BUILD IT …

THE DATA MODELING PROCESS



Collect Data

Batting, Pitching, Team, Salaries

Data Cleaning

Adjusting for Inflation, handling advanced sparse data

Feature Engineering

Removing related features

Modeling (Training)

75% Training

25% Testing

5-Fold Validation

Finalize Model

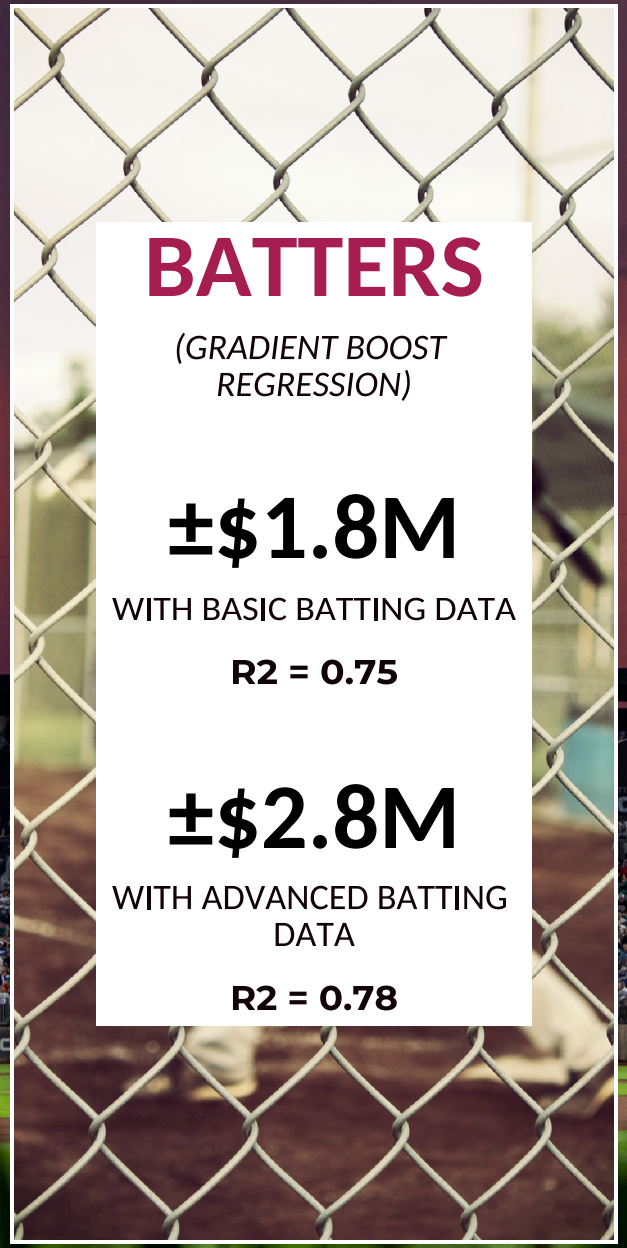
Which model(s) performed the best?

A TRIPLE PLAY

RESULTS

BASIC DATA
(2000 - 2021)

ADVANCED DATA
(2014 - 2021)



BATTERS

(GRADIENT BOOST REGRESSION)

±$1.8M

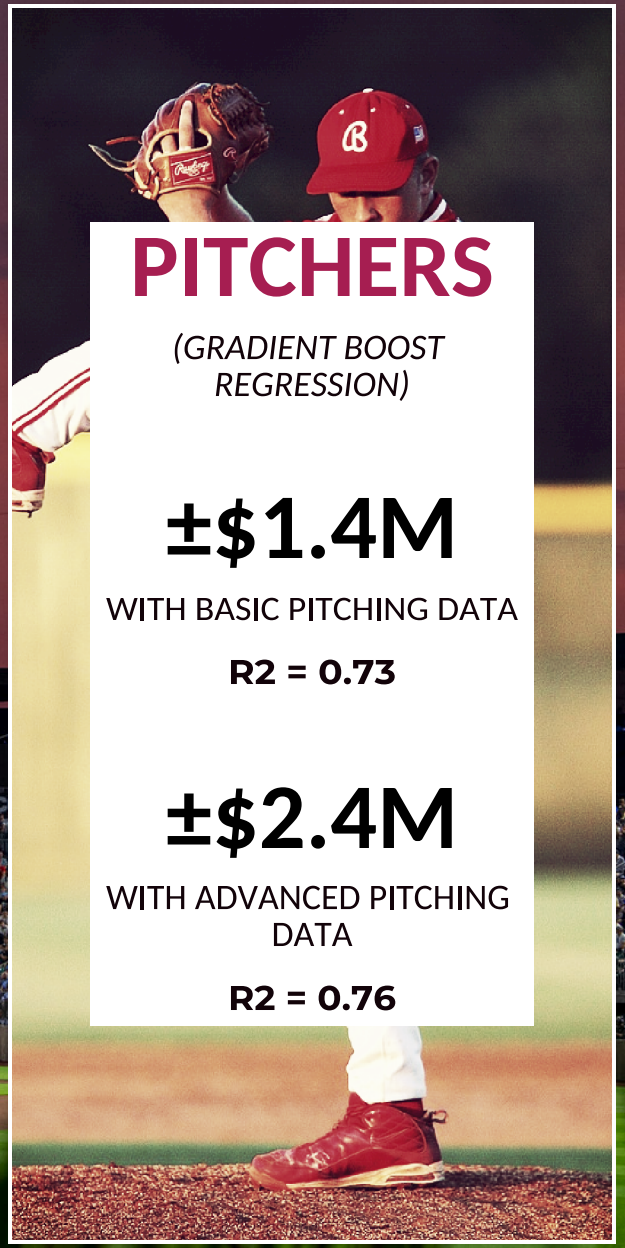
WITH BASIC BATTING DATA

R2 = 0.75

±$2.8M

WITH ADVANCED BATTING DATA

R2 = 0.78



PITCHERS

(GRADIENT BOOST REGRESSION)

±$1.4M

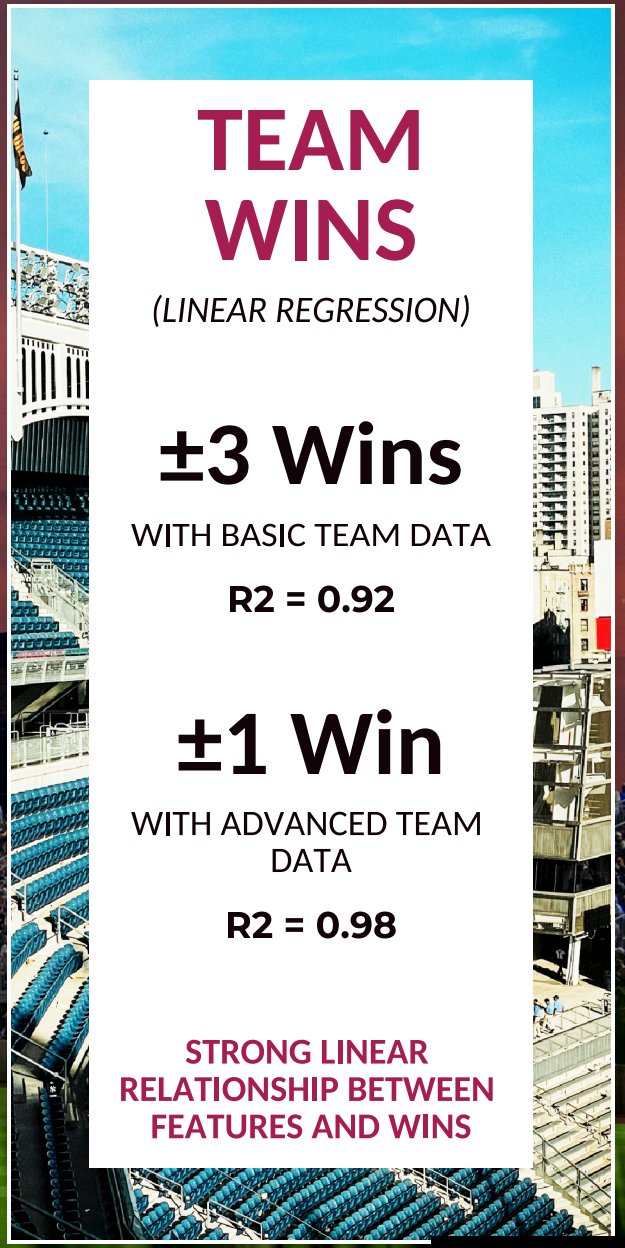
WITH BASIC PITCHING DATA

R2 = 0.73

±$2.4M

WITH ADVANCED PITCHING DATA

R2 = 0.76



TEAM WINS

(LINEAR REGRESSION)

±3 Wins

WITH BASIC TEAM DATA

R2 = 0.92

±1 Win

WITH ADVANCED TEAM DATA

R2 = 0.98

STRONG LINEAR RELATIONSHIP BETWEEN FEATURES AND WINS

CONCLUSIONS & NEXT STEPS

"Every strike brings me closer to the next home run."
- Babe Ruth

Arbitration* Value

***Arbitration: salary numbers for the upcoming season that are negotiated for players not yet eligible for free agency**

- Models do better with predicting salaries of players before free agency

Superstar Factor

- Predicting salaries of superstar players more difficult

- Marketability, Social Media Presence,

- Free Agent Market Demand

CONCLUSIONS & NEXT STEPS

“Every strike brings me closer to the next home run.” - Babe Ruth

Keys to Winning

- Build a team around players who excel at advanced metrics
- Fastballs are dominant
- Timely hits and timely pitching AKA WPA*

Win Probability Added (WPA) captures the change in Win Expectancy and credits or debits the player based on how much their action increased their team’s odds of winning

Statcast

- Advanced statistics can better explain target variable
- Relatively new technology, constant data collection

THANK YOU

Email: eric8395@gmail.com

Github: https://github.com/eric8395/baseball-analytics

LinkedIn: www.linkedin.com/in/eric-au8395