






Taking it to the Man: Using Random Forests to Beat your Bookie



Developer: Joe Kupresanin
Senior QA: Max Holiber

Associate QA: Saurabh Annadate, Michael Fedell, Finn Qiao, Surabhi Seth, Tova Simonson

TOTAL			 OPENING	 BOVADA	 interTops	 BETONLINE	 BETNOW <small>WE MAKE IT SIMPLE</small>	 CAESARS PALACE	 5Dimes	 WESTGATE LAS VEGAS	
8:15p	973	VS	Boston	8.5 o -110	8.5 o -120	8.5 o -115	8.5 o -120	8.5 o -115	8.5 o -115	8.5 o -110	8.5 o -110
	974		Kansas City	8.5 u -110	8.5 u +100	8.5 u -105	8.5 u +100	8.5 u -105	8.5 u -105	8.5 u +100	8.5 u -110
➤ Matchup ➤ Line History											

The Over / Under Bet

Bookie Says: "8.5 runs"

If the gambler thinks the game will be high scoring, he can bet "**Over**"

If the gambler thinks the game will be low scoring, she can bet "**Under**"

Bets Pay Even Money

Bet \$100, win or lose \$100

Each game is essentially a coin flip (over / under, ~50% chance for each)

The bookie charges a 10% commision, so the gambler has to win over 55% of bets to make money

Runs per Game

Typical Variation:

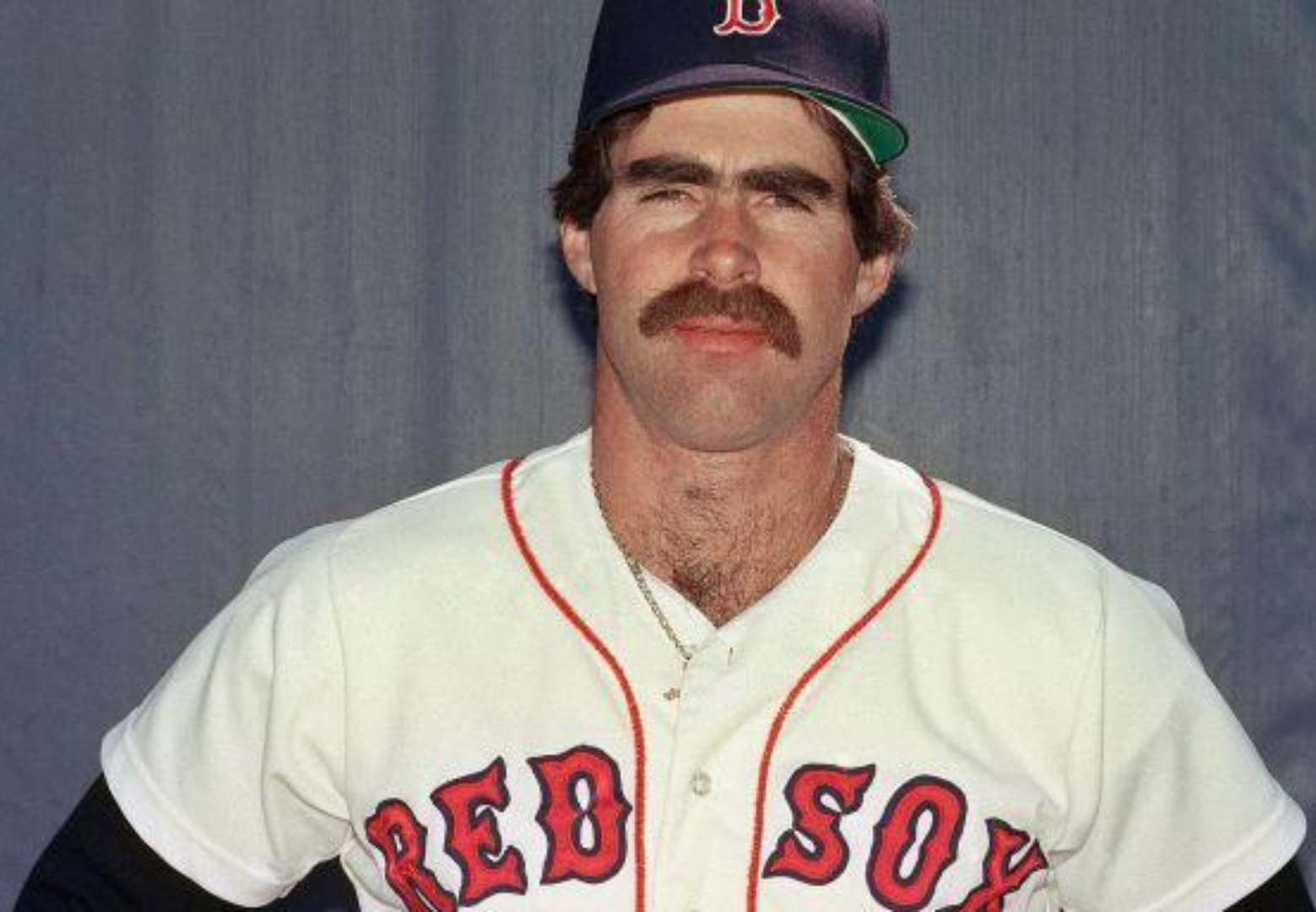
RMSE ~ 4 runs

$R^2 \sim 0.3000$

Most of the model predictions are ~8, 9, 10 runs

Motivation: Baseball Gambling is Hard*

** But in theory, one could make money*



Data:

The source code
of sport.

Leveraging the power of sports data to enhance media solutions, power sportsbook operations and to protect sports integrity.

"hit_location": 13

"hit_type": "LD"

2018 Full Season

Used for Training

- ~ 200 API pulls from <http://sportsradar.us>
- ~ 200 variables: venue, team records, weather, starting pitcher stats

Each Day

API Pull Today's Games

- ~ 15 games each day
- ~ 1-hot encode all the categorical variables from both datasets
- ~ Daily training variables are then the intersection of 2018 and today's games

Train and Gamble

Training to Predict "Runs Scored" for each game today (49 models)

- ~ The model only predicts baseball outcomes and has nothing to do with the gambling aspect of this project

Data Science Outcome:

Run today's model and enter betting lines:

Give the month as 4, 5, 6, 7, 8, or 9: 6

Give the day as 1, 2, ..., 29, 30, or 31: 3

Serving up today's model, so wait about 10 minutes, OK?

Today's random forest model parameters:

Number of Estimators: 822

Bootstrap: True

Maximum Depth: 10

Maximum Features: sqrt

Minimum Samples per Leaf: 2

Minimum Samples per Split: 7

The tuned model from 6-3

RMSE = 3.673963982449792

The tuned model from 6-3

R-Squared = 0.3439820595052312

	away	home	month	day	predicted.runs
0	LAD	ARI	6	3	9.662160
1	PHI	SD	6	3	9.736609
2	LAA	CHC	6	3	11.294068
3	HOU	SEA	6	3	9.392488

Mon June 3 [Get Today's Alerts](#)

LA Angels 0 Bot 5th
 CHI Cubs 1 2 Out

LA Dodgers -150 8:40 PM
ARI Diamondbacks T:8.5

LAD: W. Buehler (5-1, 4.03)
ARI: R. Ray (4-2, 3.59)

HOU Astros -145 9:10 PM
 SEA Mariners T:10.0

HOU: C. Martin (1-1, 5.51)
SEA: W. LeBlanc (2-2, 6.99)

PHI Phillies -145 9:10 PM
 SD Padres T:7.5

PHI: A. Nola (6-0, 4.18)
SD: E. Lauer (4-4, 4.45)

Business Outcome:

Profit / Loss Report for the 2019 season:
(Each bet is \$100)

Only bet the best pick each day: \$1100

Total Bets: 81

Winning Percentage: 57.33%

Bet every single game every day: \$-500

Total Bets: 685

Winning Percentage: 49.61%

The best team to bet on has been the San Francisco Giants

Bet only the San Francisco Giants games: \$1900

Total Bets: 46

Winning Percentage: 73.17%

Enter 1 for Today's Picks

Enter 2 for 2019 Profit Reports

Enter 3 for Daily Report

Enter 4 to Return to Main Menu

Enter 5 to Quit Program



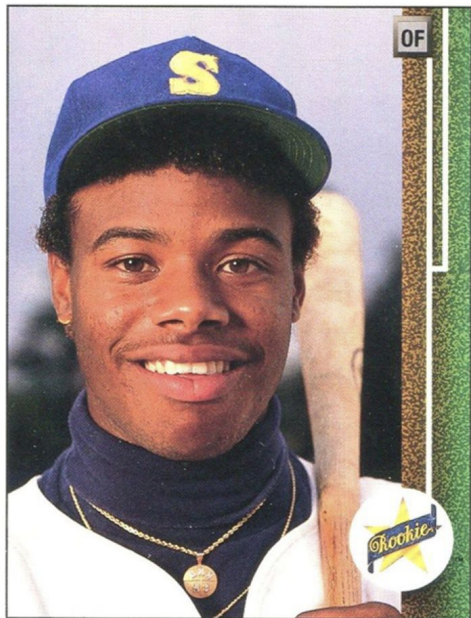
Interesting Insights:

Early Season

All Predictions High

~ First few weeks of the season, nearly all predictions exceeded 10 runs

~ Perplexing - and probably based on early season stats



Ken Griffey Jr.

Variation

RSME ~ 4.00 Runs

(backlog shot for ~1.00)

~ If baseball games with 1, 2, ..., 16, 17 runs scored are not unusual, it is remarkable that betting the top daily pick has held up through June 2nd



Thank You