DATA SCIENCE AND THE BUSINESS OF MAJOR LEAGUE BASEBALL

Matthew Horton
Josh Hamilton
Aaron Owen, PhD
DATA SCIENCE AT MLB

THERE ARE 2 DISTINCT DATA SCIENCE GROUPS, FOCUSED ON DIFFERENT ASPECTS OF THE GAME

ON THE FIELD

DATA SOURCES: Statcast AI, plate appearances, pitch tracker
AUDIENCE: Fans, clubs and players

BUSINESS / FAN FOCUSED

DATA SOURCES: Historical attendance, fan’s purchase history, app usage
AUDIENCE: Fans, clubs and internal MLB departments
WHERE DOES DATA SCIENCE FIT INTO THE ORGANIZATION AND WHO USES OUR WORK?

ORGANIZATION

MARKETING

ANALYTICS

BUSINESS / FAN FOCUSED DATA SCIENCE

GROUPS / TEAMS UTILIZING OUR WORK

- 30 TEAMS
- MLB PRODUCT TEAM
- MLB MARKETING
- MLB TICKETING
- MLB TV
- MLB SHOP
- MLB SCHEDULING / BROADCAST TEAMS
FUTURE SCHEDULE EVALUATION

CANDIDATE SCHEDULES ARE FOR 2 YEARS INTO FUTURE

CANDIDATE A

CANDIDATE B

CANDIDATE C
FUTURE SCHEDULE EVALUATION

MODEL TRAINING

7 SEASONS OF GAME DATA

day of week
month
opponent
interleague/intraleague
game time
previous attendance
previous revenue
summer vacation dates
holidays
weather variables
multi-year aggregates
feature interactions
...

LINEAR REGRESSION

SINGLE-GAME ATTENDANCE

SINGLE-GAME REVENUE

LINEAR REGRESSION
FUTURE SCHEDULE EVALUATION

CANDIDATE A

- Predicted league-wide attendance: X
- Predicted league-wide revenue: $X

CANDIDATE B

- Predicted league-wide attendance: X
- Predicted league-wide revenue: $X

CANDIDATE C

- Predicted league-wide attendance: X
- Predicted league-wide revenue: $X

COMMISSIONER’S SCHEDULING COMMITTEE’S CHOICE
<table>
<thead>
<tr>
<th>MATCHUP</th>
<th>TIME (ET)</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Diego at Cincinnati</td>
<td>12:35 PM</td>
</tr>
<tr>
<td>Seattle at Tampa Bay</td>
<td>1:10 PM</td>
</tr>
<tr>
<td>Chicago at Minnesota</td>
<td>1:10 PM</td>
</tr>
<tr>
<td>Colorado at Arizona</td>
<td>3:40 PM</td>
</tr>
<tr>
<td>Los Angeles at Texas</td>
<td>7:05 PM</td>
</tr>
<tr>
<td>KC at Baltimore</td>
<td>7:05 PM</td>
</tr>
<tr>
<td>Washington at Pittsburgh</td>
<td>7:05 PM</td>
</tr>
<tr>
<td>Cleveland at New York</td>
<td>7:10 PM</td>
</tr>
<tr>
<td>Philadelphia at Boston</td>
<td>7:10 PM</td>
</tr>
<tr>
<td>Miami at Atlanta</td>
<td>7:20 PM</td>
</tr>
<tr>
<td>Milwaukee at St. Louis</td>
<td>7:45 PM</td>
</tr>
<tr>
<td>San Francisco at Chicago</td>
<td>8:05 PM</td>
</tr>
<tr>
<td>Detroit at Houston</td>
<td>8:10 PM</td>
</tr>
<tr>
<td>New York at Oakland</td>
<td>10:00 PM</td>
</tr>
<tr>
<td>Toronto at Los Angeles</td>
<td>10:10 PM</td>
</tr>
</tbody>
</table>
SINGLE GAME TICKET DEMAND FORECASTING
SINGLE GAME TICKET DEMAND FORECASTING – MODEL

**STEP 1:**
MODEL TRAINING

- 3 PREVIOUS SEASONS OF GAME DATA
  - number of days before game
  - day of week
  - month
  - opponent
  - interleague/intraleague
  - promo/no promo
  - ...

- GRADIENT BOOSTED REGRESSOR

- NUMBER OF TICKETS SOLD

**STEP 2:**
PREDICTIONS

- CURRENT TICKET SALES FOR ALL GAMES

- TRAINED MODEL

- GAME 29

- TICKET SALES

- DAYS BEFORE GAME

- ACTUAL

- PREDICTED
PROMOTION SCHEDULE OPTIMIZATION - MODEL

**STEP 1:**
MODEL TRAINING

3 PREVIOUS SEASONS OF GAME DATA
- series start
- day/night
- day of week
- month
- opponent
- promo type

GRADIENT BOOSTED REGRESSOR

REVENUE PREDICTION BASED ON CURRENT PROMOTION SCHEDULE

**STEP 2:**
MONTE CARLO SIMULATION

PROMOTION SCHEDULE RANDOMIZED x10,000

TRAINED MODEL

10,000 REVENUE PREDICTIONS
PROMOTION SCHEDULE OPTIMIZATION - RECOMMENDATION
PROMOTION SCHEDULE OPTIMIZATION - RECOMMENDATION

HYPOTHETICAL REVENUE

ORIGINAL PROMOTION SCHEDULE

NO PROMOTIONS

WORST 10% OF SIMULATED PROMOTION SCHEDULES

BEST 10% OF SIMULATED PROMOTION SCHEDULES

1. **FIREWORKS**
   - Original schedule x 2
   - Original schedule x 3
   - Original schedule x 2
   - Original schedule x 1
   - Original schedule x 3

2. **SHIRT OR CAP**
   - Original schedule x 2
   - Original schedule x 3
   - Original schedule x 2
   - Original schedule x 1
   - Original schedule x 3

3. **BOBBLEHEAD - OR - FIGURINE**
   - Original schedule x 1
   - Original schedule x 5
   - Original schedule x 1
   - Original schedule x 1
   - Original schedule x 1

SIMULATIONS

**NO PROMOTIONS**
TEAM AVIDITY METRIC

Strong Mets Fan

FAN SEGMENTATION

Team Fan

LIFETIME VALUE

Ticketing LTV: $500
Shop LTV: $100
MLB.tv LTV: $100
Overall LTV: $700

PLAYER AVIDITY

Jacob DeGrom
Pete Alonso
Mike Trout
TEAM AVIDITY – DEVELOPMENT

6 PREVIOUS YEARS OF FAN DATA

- Explicit Signals
  - Email opt-ins
  - Ballpark app
  - MLB.TV streams
  - Ticket scans
  - Shop purchases
  - Ticket purchases

- Engagement

- Share of Fan’s Spend

DATA SOURCE

FEATURES

TEAM AVIDITY

TeamAvidity(fan, team) = ExplicitSignals × W_{ES} + Engagement × W_{E} + Spend × W_{S}

SCORE AND RANK

STANDARDIZE AND SEGMENT
TEAM AVIDITY – USE CASES

IDENTIFY AND TARGET OUT-OF-MARKET FANS

HOME FIELD ADVANTAGE

OFTEN MOST PREDICTIVE FEATURE IN MODELS
FAN SEGMENTATION – MODEL

3 PREVIOUS YEARS OF FAN DATA

DATA SOURCE

- MLB.TV
  - Attended Games
- participating teams per game

FEATURES

NETWORK ANALYSIS

- Degree centrality,
  - Clustering coefficient

MLB.TV

FEATURES

- Attended Games

ALL FANS

- ROOKIES
  - Casual or New Fan
- TEAM FANS
  - Mostly Interested in a Single Team
- VETERANS
  - Interested in Many Teams
FAN SEGMENTATION – USE CASES

ROOKIES

TEAM FANS

VETERANS
FAN LIFETIME VALUE (LTV) – MODEL

3 PREVIOUS YEARS OF FAN DATA

Business Lines

Features

ticket spend
total num tickets
unused tickets
ticket resell ROI
...  
shop spend
shop returns
shop unique products
...  
MLB.tv total mins watched
MLB.tv subscriber type
MLB.tv num cancels
MLB.TV num year subscriber
...

REPURCHASE Model
(Gradient Boosted Classifier)

Probability of Repurchase

[0, 1]

Predicted LTV

Potential Spend Model*
(Gradient Boosted Regressor)

Predicted Potential Spend

[0, ∞)

*only trained on fans that went on to spend again
FAN LIFETIME VALUE (LTV) – USE CASES

EACH MLB FAN

MLB shop.com

MLB.tv

TOTAL LTV

MLB.tv LTV

SHOP LTV

TICKETING LTV

LTV SEGMENTATION

POTENTIAL SPEND

PROBABILITY OF REPURCHASE

EVALUATING MARKETING/ADVERTISING CAMPAIGN EFFICACY

A/B TESTING

RESPONSE

Watch These Games On Demand

Full Archive

A vs. Astros June 2

B vs. Tigers June 3

C vs. White Sox June 9

vs. Dodgers June 3

vs. Braves June 9

vs. White Sox June 9

Watch >
PLAYER AVIDITY – MODEL

CURRENT FAN + PLAYER DATA

- fan’s team avidity
- fan’s location
- player’s MLB popularity
- player’s team popularity
- player’s performance

LATENT VARIABLE MODEL
EXPECTATION-MAXIMIZATION ALGORITHM

PREDICTED
- All-Star Votes
- Shop Sales
- Website Views

OBSERVED
- All-Star Votes
- Shop Sales
- Website Views

FAN-PLAYER AVIDITY
CUSTOMIZED FAN CONTENT

IMPACT OF ROSTER CHANGES ON A CLUB’S FANBASE
TEAM AVIDITY METRIC

FAN SEGMENTATION

LIFETIME VALUE

Ticketing LTV: $500
Shop LTV: $100
MLB.tv LTV: $100
Overall LTV: $700

PLAYEER AVIDITY

Jacob DeGrom
Pete Alonso
Mike Trout

TICKET PACKAGE RENEWAL LEAD SCORE

Top 20% of fans likely to renew

SEASON TICKET HOLDER RISK

Not a season ticket holder

MLB.TV ENGAGEMENT CAMPAIGN

Moderately engaged
Received email last week

TARGETED TICKET GUIDE

Received notification about CLE vs. NYM series
OTHER FAN-BASED MODELING

LEAD SCORING

SEASON TICKET HOLDER RISK SCORE

MLB.TV ENGAGEMENT

TARGETED TICKET GUIDE
QUICK SUMMARY

• OVERVIEW OF DATA SCIENCE AT MLB
• METRIC INTRODUCTIONS
  • GAME
  • FAN

BETTER SERVE THE 30 CLUBS & OUR MILLIONS OF FANS
Rate today’s session

We’re living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you’re often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we’re uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.
QUESTIONS?

• DataScience@mlb.com

• OPEN POSITIONS: www.mlb.com/jobs

• TECH BLOG: https://technology.mlblogs.com/