Predicting Major League Baseball Batting Averages Using Machine Learning and Statcast Data

Joey Capps

What is a batting average?

BA = H / AB

Where:

H = hits

AB = at bats

- Only counts as an AB if something other than HBP, walk, or sacrifice hit happens

Who cares about batting avg?

Fans

- How good will my favorite player do?

General managers

- How much do we pay this guy?

Fantasy players

- Who do I draft?

Gamblers

- Who do I bet on?

Objective

Given any previous stats, what will player X’s batting average be over the span of a season?

- Success will be measured using mean absolute error (MAE)
- The lower the better
- Which stats should we grab to achieve this?

Data collection

Competing models usually use features in 3 categories:

1. Standard data
   - HR, AB, BA, age, etc.
2. Advanced data
   - BABIP, ISO, etc.
3. Statcast data
   - xBA, speed, etc.
   - 2015+ only

$$\text{BABIP} = \frac{H - HR}{AB - K - HR + SF}$$

Competition

- Limitation is that obtaining raw prediction data is sometimes **paywalled,** and implementation details area almost always **proprietary.**
- Article from 2025 compared 14 different models' accuracy from 2024
- Steamer performed 6th in hitting, 8th in batting average
- The Bat X performed 1st in both hitting and batting average
- Use Wayback machine to compare our model with these

Feature engineering

1. Collect standard and Statcast data, compute advanced data
2. Merge these all together via inner joins
3. Use lagged columns for some number of years (e.g. BA_lag1, BA_lag2, BA_lag3)
   - 3 seemed to work best
4. Use cumulative career stats (e.g. CareerAvg_BA, CareerAvg_xBA, CareerAvg_BBE)
5. Excluded rookies
   - No previous year data was collected
6. Big column size of 236 - use PCA

Final dataset was 1,845 from 2015 - 2025

Best predictors?

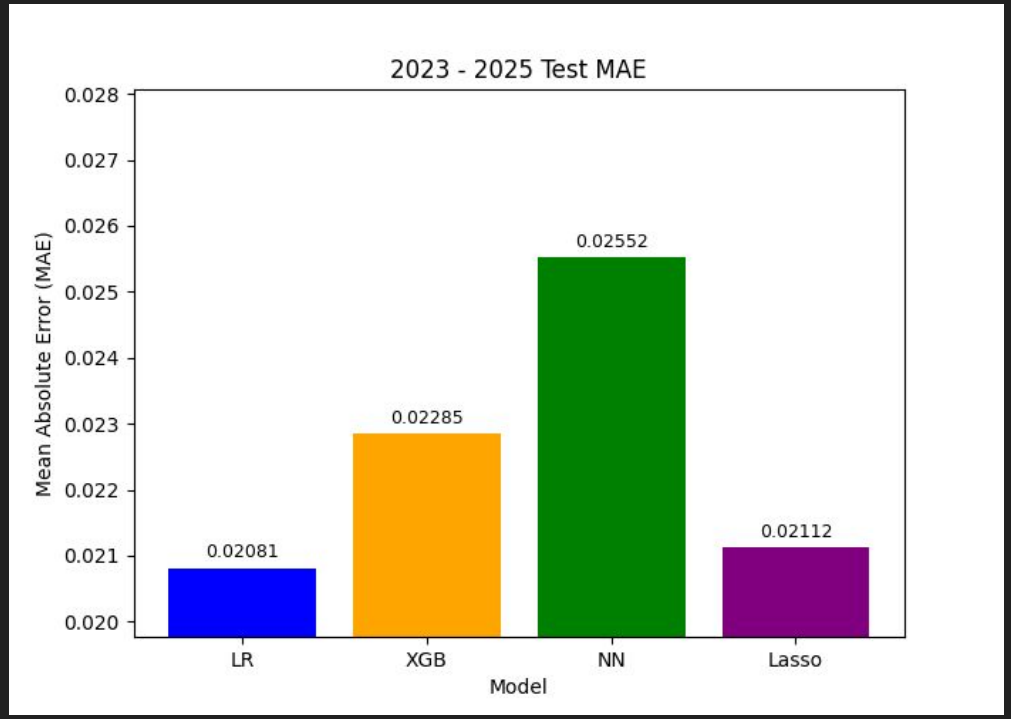
Measured using F-score

- Career avg beats normal lag
- Est_bat (xBA) beats normal BA

Feature Name	F-Score
CareerAvg_est_ba	292.32
CareerAvg_BA	286.4
est_ba_lag1	243.1
BA_lag1	215.63
est_ba_lag2	191.31
est_ba_lag3	181.09
BA_lag2	162.9
BA_lag3	160.77
bbe_lag2	134.56
bbe_lag1	134.56

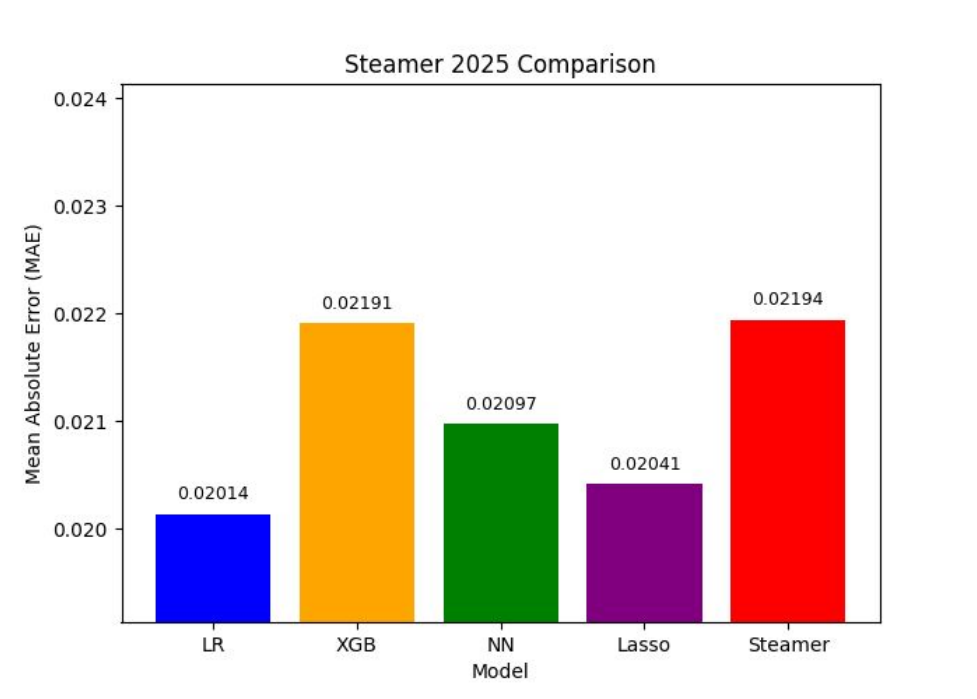
Results - 2023 to 2025 test set

- Test size 971



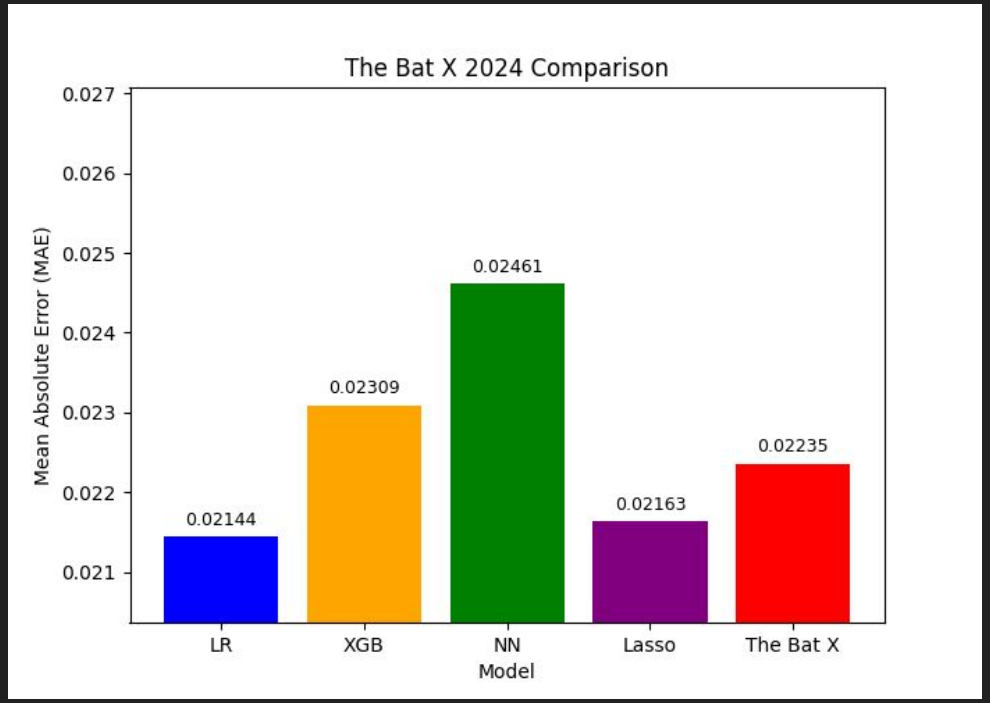
Results - 2025 Steamer

- Test size 360



Results - 2024 The Bat X

- Test size 336



Results - what next?

1. Showed that this strategy can beat the best models
2. Future work involves expanding this strategy to other stats and building a comprehensive model
3. Need to test on other competitors over wider time frame to be sure
4. Including pre-2015 stats may further decrease MAE