Predicting Division III Softball Outcomes

By: Malcolm Ehlers, Gustavo Garcia-Franceschini, Lawrence Jang, Bin Zheng
Advisor: Ronald Yurko

BACKGROUND

- CMU's Division 3 Softball team was founded in 2019
- We have collected practice data, play-by-play data, and batter & pitcher statistics
- Our goals are:
  - **Analyze player practice performance and explore the relationship between practices and games**
  - **Model softball outcome probabilities from the play-by-play data and batter & pitcher statistics**
- This will allow Coach Monica Harrison to plan practices and have an additional tool to use for strategizing and deciding lineups

DATA

Practice data:

- 90 observations, 20 variables. Each observation is a player-season
- We found no meaningful results in the practice data
- Very little data, the team is young and the 2020 and 2021 seasons were heavily affected by COVID-19

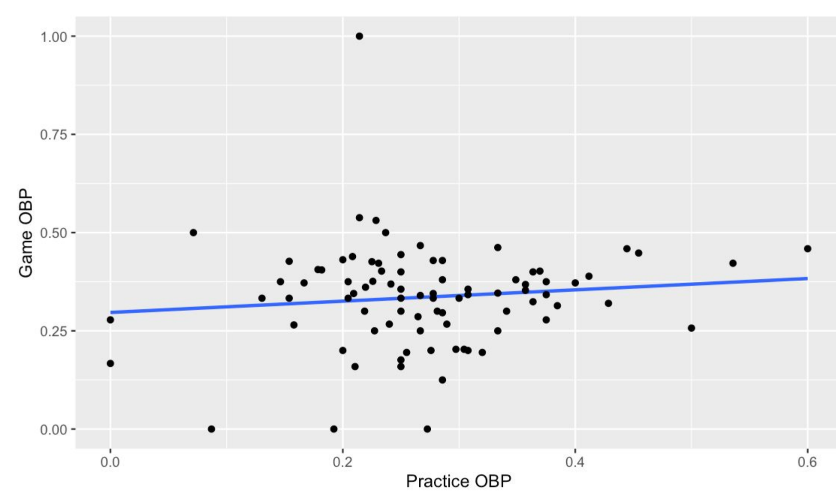


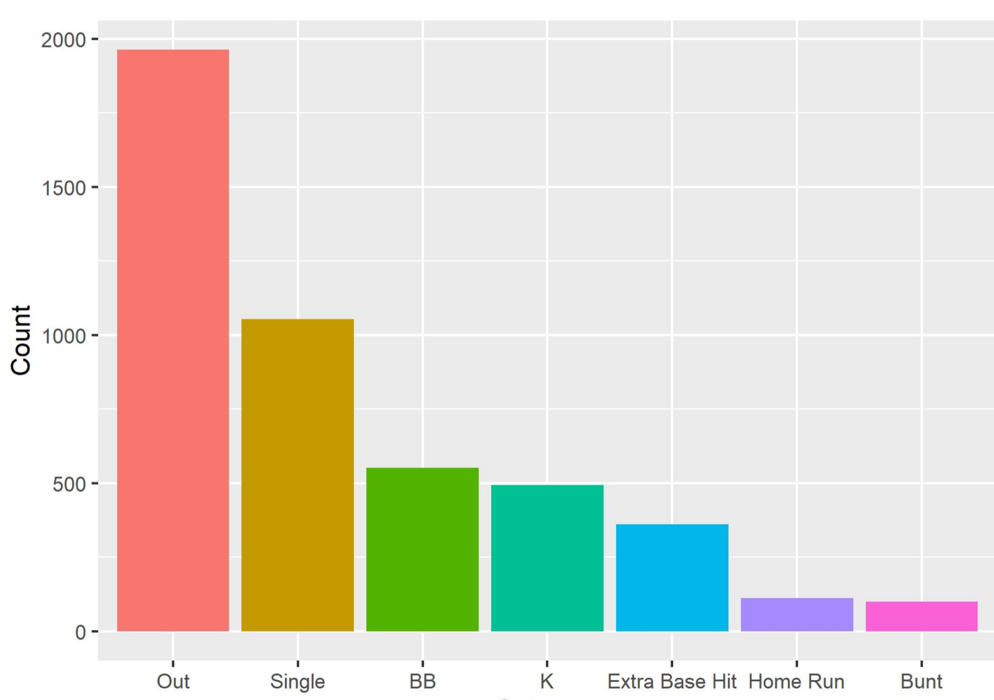
Figure 1: Scatter plot of on-base percentage (OBP) in competitive games vs. in practice. There is no significant relationship between the two (p = .61)

Processed data:

- Merged play-by-play data with batter & pitcher statistics (2022 season only)
  - Player level batter statistics, school level pitcher statistics
- 4637 observations, 91 variables. Each observation is an at-bat
- Spans across 124 games, played by 17 schools within CMU's schedule

Predictor variables: We utilized 16 predictor variables for modelling. By the nature of the merged data, the predictor variables can be categorized as follows:

Play-by-play	Batter statistics	Pitcher statistics
Innings Base-out scenarios	Strikeout percentage (K%) On-base percentage (OBP)	Batting average against (BAA) Avg. OBP against

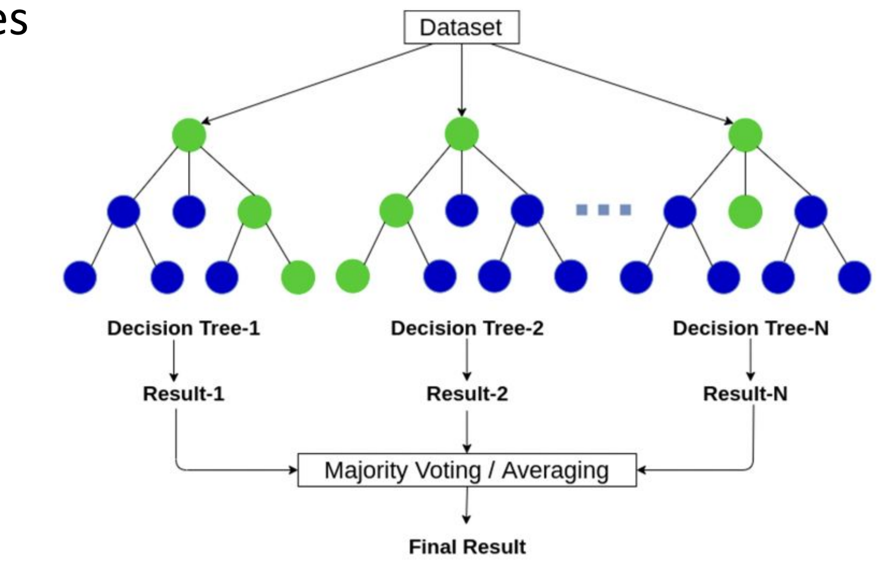


Response variable: Seven at-bat outcomes: Out, Single, BB (walk), K (strike), Extra Base Hit (EBH), Home Run (HR), and Bunt

Figure 2: Distribution of outcomes. Outs and singles are the two most common outcomes

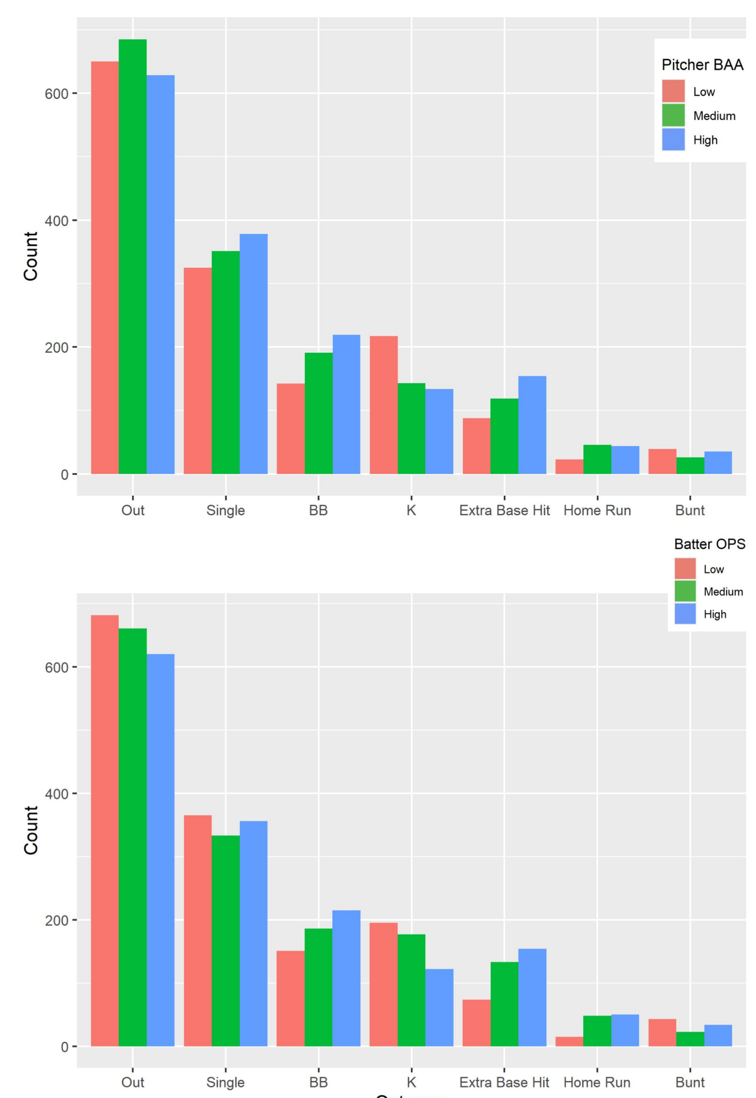
METHODS

- Built multinomial logistic regression and random probability forest models to predict outcome probabilities
  - Multinomial Logistic Regression: $\log\left(\frac{p(m|x)}{p(Out|x)}\right) = X\beta_m, \quad m \in \{Single, BB, K, EBH, HR, Bunt\}$
  - Random Probability Forest: $\hat{p}(outcome|x) = \frac{1}{B}\sum_{b=1}^{B}\hat{p}_b(outcome|x)$
- To avoid data linkage, we assigned games (rather than individual at-bats) to cross validation folds

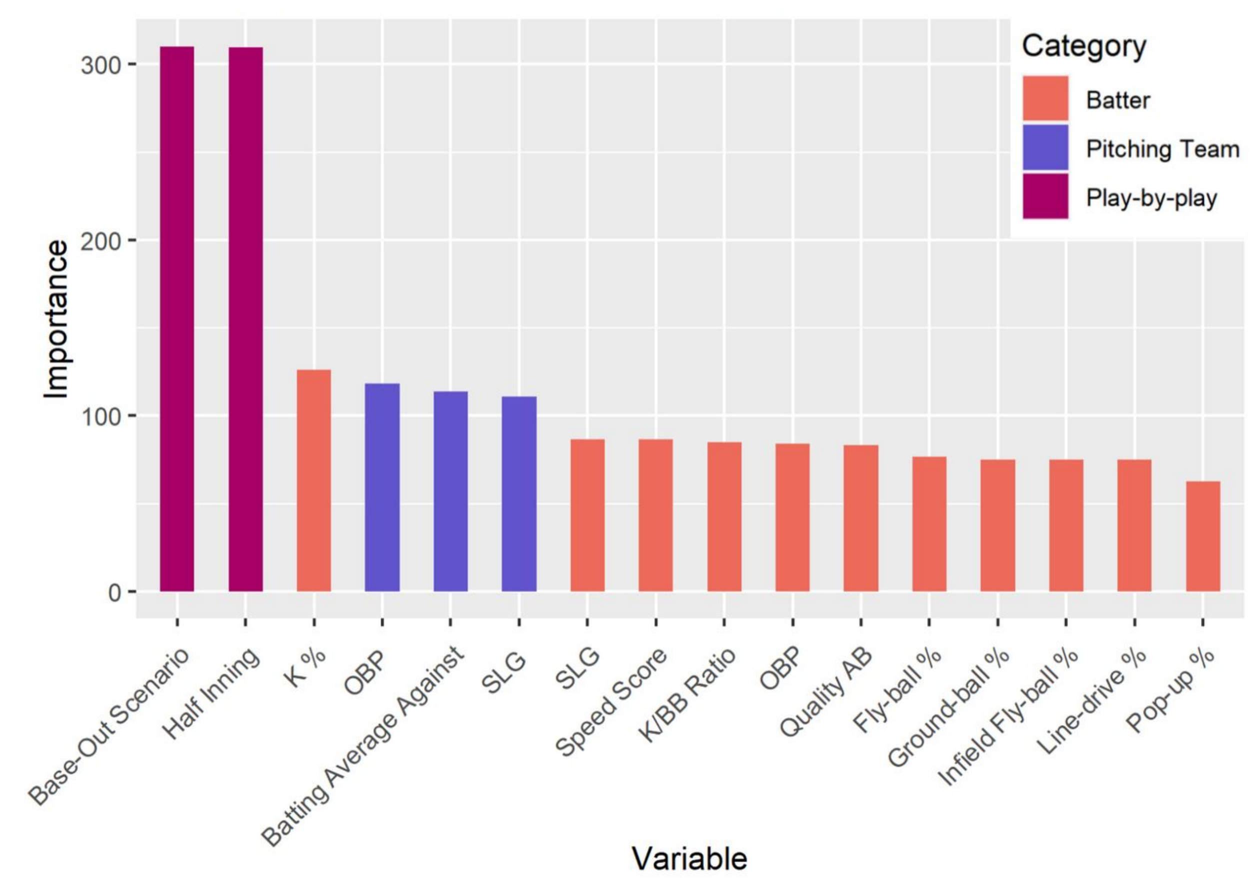


ANALYSIS & RESULTS

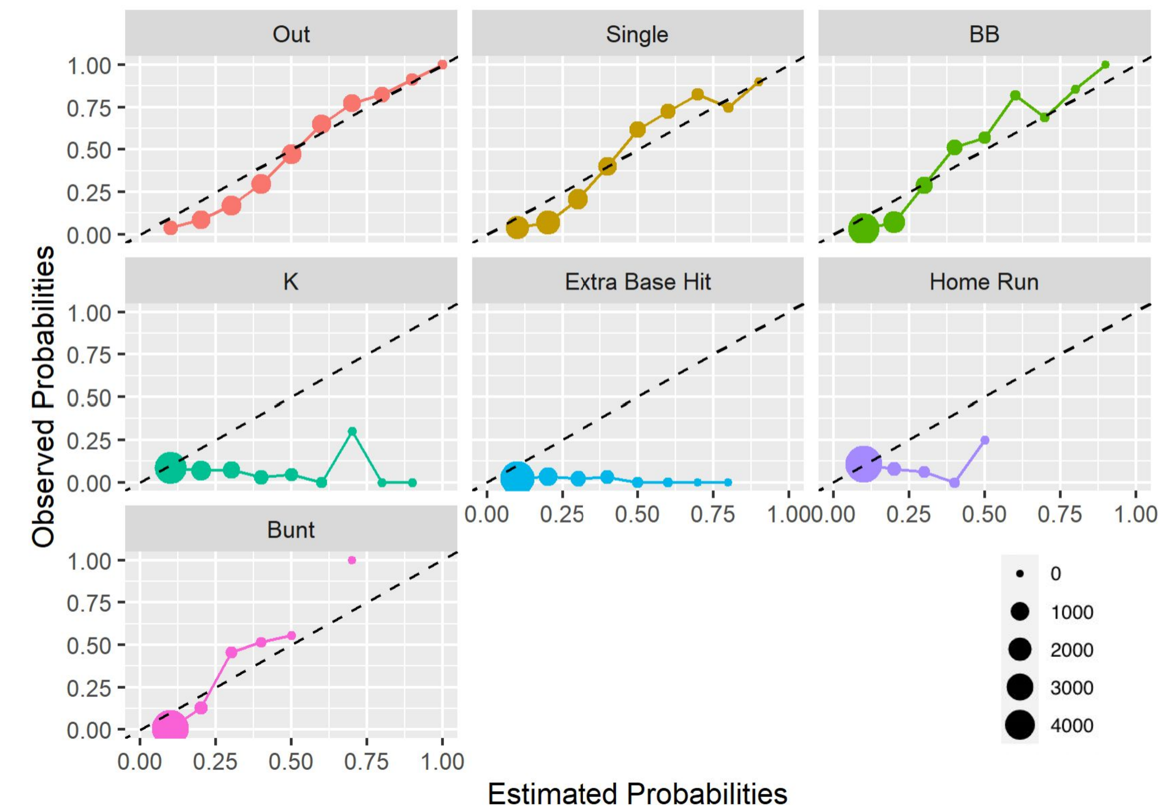
Distribution of outcome varies by batter & pitcher statistics



Our best model is the random probability forest. Play-by-play variables are the most important for our model



Random probability forest model predicts Out, Single, BB (Walk + HBPSs), and Bunts well



Player	Out	BB	Bunt	EBH	HR	K	Single
Abby	45.41%	7.75%	1.00%	2.49%	4.34%	5.21%	33.81%
Becky	50.72%	4.05%	0.36%	1.49%	0.68%	1.69%	41.01%
Carla	63.58%	5.49%	0.25%	8.30%	0.83%	3.12%	18.43%

Player	Out	BB	Bunt	EBH	HR	K	Single
Abby	43.15%	27.55%	0.69%	5.23%	1.40%	12.96%	9.04%
Becky	45.28%	4.53%	0.65%	3.07%	1.54%	7.78%	37.16%
Carla	24.57%	3.04%	1.14%	19.07%	0.59%	26.72%	24.87%

Sample batter probability outcome matrices. Left table is against Emory (highest BAA), right table is against Trine(lowest BAA)

CONCLUSION

- Did not find any meaningful results in the practice data
  - Lack of data with newly founded team and several seasons affected by COVID-19
- We fit a random probability forest model to predict softball outcomes from play-by-play data and batter & pitcher statistics
  - Best at predicting outs, singles, walks, and bunts
  - Not very good for strikes, extra base hits, and home runs
- Next steps: Collect more data and create RShiny app for better user experience

REFERENCES

6–4–3 charts. 6. (2022, November). Retrieved March 14, 2023, from http://643charts.com/ [Dataset]

Christy, M., Ohl, Z., Willis, A., and Zeng, E. (2022, May). *Varsity Softball Capstone*. [Scholarly project].

Powers, S., Hastie, T., and Tibshirani, R. "Nuclear penalized multinomial regression with an application to predicting at bat outcomes in baseball." *Statistical modelling* 18.5-6 (2018): 388-410.