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

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:

<table>
<thead>
<tr>
<th>Play-by-play</th>
<th>Batter statistics</th>
<th>Pitcher statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Innings</td>
<td>Strikeout percentage (K%)</td>
<td>Batting average against (BAA)</td>
</tr>
<tr>
<td>Base-out scenarios</td>
<td>On-base percentage (OBP)</td>
<td>Avg. OBP against</td>
</tr>
</tbody>
</table>

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

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)

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: \[ \log \left( \frac{p(m|x)}{p(\text{Out}|x)} \right) = X\beta_m, \ m \in \{\text{Single}, \text{BB}, \text{K}, \text{EBH}, \text{HR}, \text{Bunt}\} \]
  - Random Probability Forest: \[ p(\text{outcome}|x) = \frac{1}{B} \sum_{b=1}^{B} p_b(\text{outcome}|x) \]
  - 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

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