Omar Badawi, Charles Bloss, Thomas Emerick, Sebastian Escobar-Mesa, Connor Klebrowski

Our project started by trying to study and understand the relationship between how a softball is hit and what happens afterward. To be able to do this, we began with analyzing MLB Statcast data, which contains millions of batted ball data. This helped us explore broad patterns regarding these batted balls and look at factors such as exit velocity, launch angle, and offensive success. We created several visualizations, such as heatmaps and scatter plots, which helped us understand the values for exit velocity and launch angle at which most hits occur. Starting with the MLB data also gave us an opportunity to be able to practice building the tools we would later need for the LSU Softball data. This would include data cleanup steps and producing early versions of our models, so that they can be fine-tuned later on.

Once we got access to the softball data, we were able to shift our focus entirely to the softball environment. This started with cleaning the new dataset, since not every play in the softball data included the variables we needed for what we were aiming to study. After filtering down to only the usable batted balls and simplifying some of the outcome categories, we began building our models. The models were built to estimate how likely each different type of result was based on how the ball was hit. In order to do this, we used a k-nearest neighbors (kNN) model. This model works by comparing each batted ball to similar ones from the past. With this model, we made our first outcome probability maps with the LSU data. This was able to help us identify which combinations of different exit velocity and launch angle were most beneficial for the hitters.

Building off these results, we add some smoothing steps using a Generalized Additive Model to create more clean outcome probabilities by having continuous surfaces rather than

scattered and noisy estimates. These smoothed surfaces, especially the ones dealing with xwOBA made it much easier to find patterns. The visuals we were able to create from all of this clearly show how just how much the quality of contact affects the outcome. Together, the kNN model and the smoothed surfaces from the GAM allowed us to create expected statistics for xBA and xwOBA with the LSU softball data. These are widely used stats in softball and baseball analytics.

To convert the outcome probability information into useful performance metrics, we constructed expected batting average (xBA) and expected weighted on-base average (xwOBA) using the modeled outcome probabilities. Expected batting average represents the total probability that a batted ball will result in any type of hit. Expected weighted on-base average extends this idea by assigning different weights to each type of outcome based on its in-game value, giving greater importance to extra-base hits and home runs. These metrics provide a more stable and informative evaluation of hitter performance than traditional results-based statistics. The expected values produced by our models represent broad offensive trends across exit velocity and launch angle. High exit velocities combined with optimal launch angles produce the highest expected offensive value, while low exit velocities or extreme launch angles produce lower expected outcomes. These trends were consistent with what we observed in our MLB data analysis, which reinforced our original beliefs that the contact quality would behave similarly in professional baseball and collegiate softball.

The kNN model performed well at classifying batted-ball outcomes, particularly for outs and home runs. These outcomes showed strong separation in exit velocity and launch angle space, making them easier to classify accurately. Singles and extra-base hits showed more overlap, which is expected because there are some other factors at play such as the runner's

speed or the ball location, which could impact the runner's decision to attempt to reach another base or not. Overall, the kNN model provided a reliable local classification framework for understanding how individual batted balls translate to offensive results. The GAM model achieved strong predictive accuracy when estimating xwOBA across the dataset. The smoothed surface produced by the GAM explained the majority of the variation in expected offensive value and aligned closely with the observed trends in the data. This also provided a visually interpretable representation of how exit velocity and launch angle interact to influence run value.

While we are very satisfied with our results, there are some limitations that must be acknowledged. The first being that the dataset was not nearly as large as the MLB dataset, which led to some outliers, due to inclement weather, being more prominent than we would have hoped. The kNN model, in particular, is more sensitive to sample density and outliers, but we expect that the already strong results will only improve once we have more years of data, as well as the data from other schools.

Future groups working on this project should take advantage of the analysis of other factors of a batted ball. Trackman, the technology used to capture the ball data, collects a lot of information on each hit, and perhaps there are factors other than exit velocity that show promise for estimating hit quality. There are some other ideas to incorporate once we have more data, such as situational context, like pitch type, count, and defensive alignment, if that data becomes available. The expected-statistic framework could also be adapted to evaluate pitchers by modeling how they suppress quality contact in EV–LA space.

Overall, this project demonstrated a unique opportunity that not many have. We were able to demonstrate how mathematical modeling and data analysis can be directly applied to real-world decision making in college athletics. It showed that exit velocity and launch angle

alone can explain a large portion of offensive performance and that expected statistics provide a more stable measure of hitter ability than traditional results-based metrics. The project also emphasized the importance of careful data cleaning, visualization, and model validation in applied sports analytics, and we are grateful for the opportunity to take part in this project.