Expected Value Modeling of Batted-Ball Outcomes in LSU Softball

MATH 4020: Final Presentation

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

Louisiana State University

Research supported by the
Mathematics Department @ LSU

December 2, 2025

Project Objectives

- Analyze the relationship between exit velocity (EV) and launch angle (LA) on batted ball outcomes using probabilistic classification methods.
- Construct LSU-specific xBA and xwOBA as functions of EV and LA.
- Estimate local outcome probabilities in EV-LA space using kNN classification.
- Regularize and smooth discrete probability estimates using Generalized Additive Models (GAM) to obtain continuous outcome surfaces.
- Develop interpretable visualizations of expected outcomes across EV-LA space for applied use in hitter evaluation and player development.

Midterm Overview

- We analyzed MLB Statcast data to study relationships between contact quality and outcomes.
- We created two different models
  - Exit Velocity vs Launch Angle (based on xBA)
  - Where the hits live (with respect to outs, base hits, and home runs)
- Defined wOBA (weighted On Base Average) and xwOBA (expected weighted on-base average)
- Set up goals for ourselves for the second half of the semester
  - Building a kNN model utilizing softball data once available.
  - Compare LSU Softball data to MLB data.

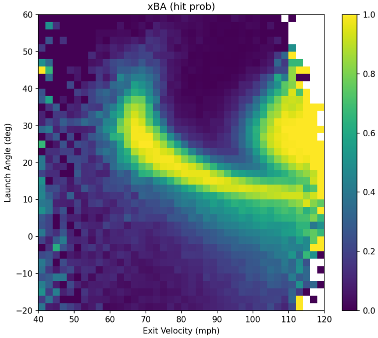
Data Cleaning Pipeline

- We began with MLB Statcast data from 2020–2024, which includes every pitch and batted ball.
- From that dataset, we kept only the in-play events that had both a valid exit velocity (EV) and launch angle (LA) recorded.
- We then mapped outcomes to a simple numeric scale for easier analysis:

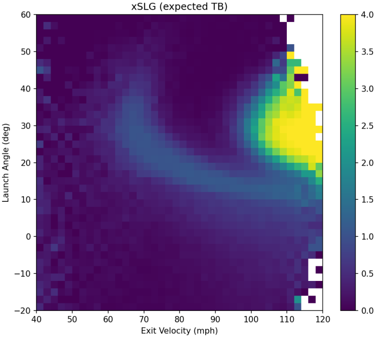
$$\text{Out} = 0, \quad 1\text{B} = 1, \quad 2\text{B} = 2, \quad 3\text{B} = 3, \quad \text{HR} = 4.$$

- The cleaned dataset now focuses only on the variables that matter most: who hit the ball, how hard it was hit, how high it was hit, and what the result was.
- We then, for the final project, applied these principles to the LSU Softball data to construct our models.

MLB Data Key Visuals: Exit Velocity vs. Launch Angle



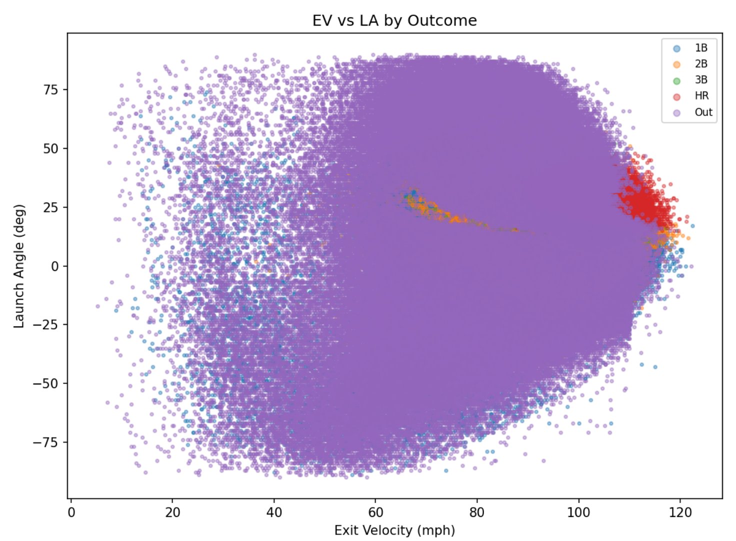
Average Hit Probability by EV–LA Zone



Average Total Bases by EV–LA Zone

Each heatmap divides exit velocity and launch angle into bins and shows how often balls in that zone became hits or extra-base hits. The bright areas represent the "sweet spot" for contact quality.

Where Do Hits Live?

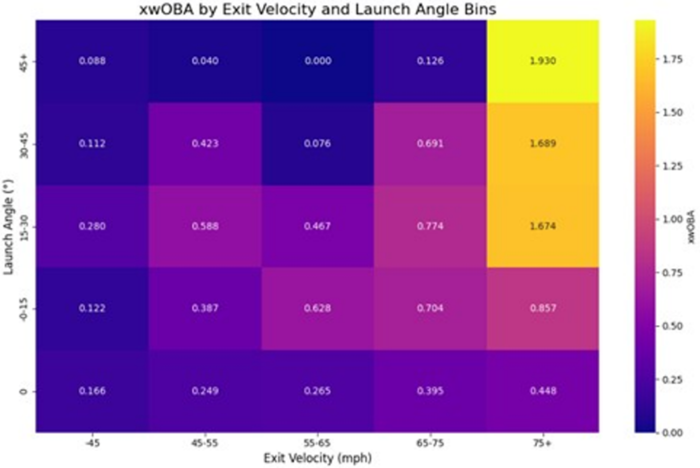


Each point represents one batted ball, plotted by its exit velocity and launch angle.

Progress Since the Midterm

- **Data shift:** Moved from public MLB Statcast data to LSU Softball Trackman data provided by the softball staff.
- **Pipeline adaptation:** Reused the EV–LA framework from the midterm and adapted it to the structure, scale, and distribution of LSU Softball batted-ball data.
- **kNN modeling:** Trained a k-nearest neighbors (kNN) classifier on EV–LA to estimate local outcome probabilities (out, single, XBH, HR) for each contact point.
- **GAM smoothing:** Fit a generalized additive model (GAM) to smooth the local probabilities into a continuous xwOBA surface over EV–LA space.
- **LSU-specific expected stats:** Used the kNN $\rightarrow$ GAM pipeline to construct LSU-specific xBA and xwOBA and generate the visualizations shown in the following slides.

Binned xwOBA Surface

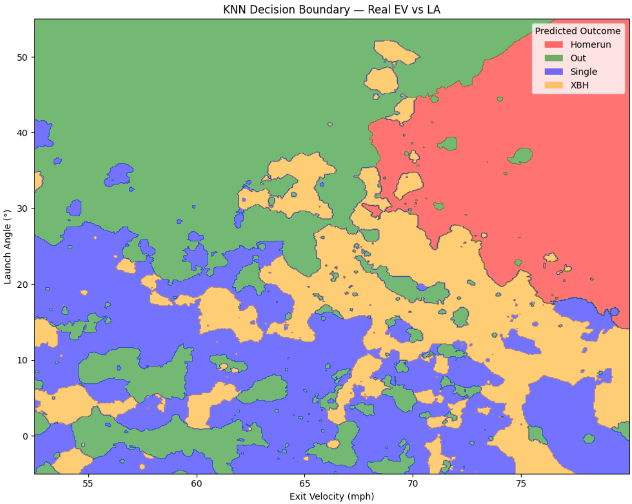


Binned LSU xwOBA values across exit velocity and launch angle groups, illustrating how contact quality influences expected offensive value.

k-Nearest Neighbors (kNN)

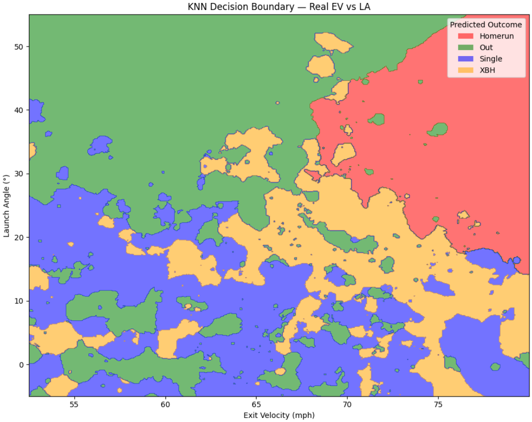
- kNN predicts an outcome based on the k most similar data points.
- For a new batted ball:
  1. Find nearby hits in EV - LA space.
  2. Average their outcomes to estimate hit probability.
- Example: a ball hit at **75 mph** and $\mathbf{30}^{\circ}$ is compared to a similar past contact to estimate its chance of being a hit.

kNN Decision Boundary



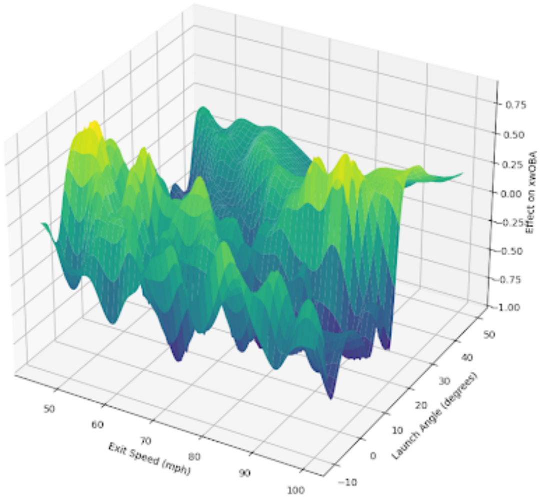
Local outcome classification from k-nearest neighbors, showing dominant batted-ball outcomes (out, single, extra-base hit (XBH), home run (HR)) by contact profile.

kNN Model Performance



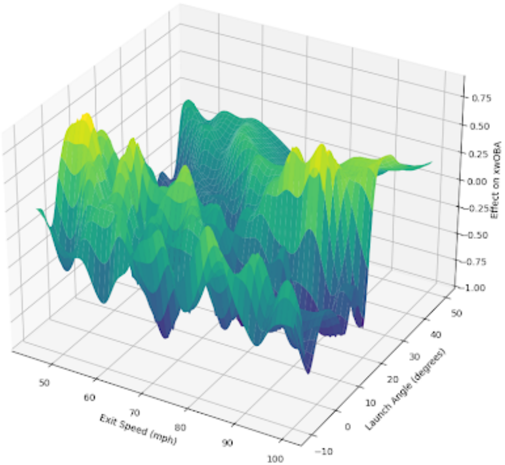
- The kNN model predicts categorical batted-ball outcomes.
- Mean cross validated F1 score $= 0.831 \pm 0.020$.
- Most accurate on outs and home runs.
- Complements the GAM well by providing discrete outcome predictions.

GAM Visualization



Smoothed EVLA interaction surface from the GAM showing the combined effect of EV and LA on xwOBA.

GAM Model Performance



- The GAM predicts xwOBA for each batted ball.
- Strong accuracy:
  RMSE $= 0.163$
  MAE $= 0.099$
  $R^2 = 0.944$
- Predicted xwOBA aligns closely with observed values.
- Produces a smooth continuous expected value model across EV–LA space.

Moving Forward

- Integrate model outputs into hitter evaluation and player development workflows.
- Validate model stability using additional seasons of LSU Softball Trackman data.
- Extend the framework to situational expected-value modeling, including context such as pitch type and game situation.
- Explore player-specific calibration and opponent-adjusted expected statistics.
- Incorporate weather-adjusted corrections to account for environmental distortions.

Acknowledgments

- We thank Dr. Zach Jermain and the LSU Softball staff for the project opportunity and their insights.
- We thank Dr. Nadejda Drenska and Dr. Peter Wolenski for their guidance on this project.
- We also thank Fernando Heidercheidt, Maganizo Kapita, and Matthew Lemoine for their assistance throughout this work.

References

T. Petersen, "Augmenting statcast expected batting average (xBA) with sprint speed," *MLB Technology Blog*, 2019.

Published September 21, 2019.

T. Nestico, "Modelling xwOBA (with KNN)," *Medium: Thomas Nestico*, 2024.

Published February 14, 2024.

W. Zhao, V. S. Akella, S. Yang, and X. Luo, "Machine learning in baseball analytics: Sabermetrics and beyond," *Information (MDPI)*, vol. 16, no. 5, 2025.