Utilizing Graph Neural Networks to Model Major League Baseball At-Bat Outcomes

CSDS 446: Final Presentation

Selah Dean

Case Western Reserve University

April 23, 2025

Table of Contents

Machine Learning in Sports

Machine Learning in Sports

- Koseler & Stephan (2017) performed systematic review of machine learning applications in baseball.
  - Primarily used traditional machine learning algorithms such as SVM and KNN.
- Silver (2020) developed a neural network model named Singlearity-PA that is designed to predict the outcome of plate appearances in MLB.
  - Provides more accurate predictions compared to traditional methods.

Graph-Based Machine Learning in Sports

- Tracy et al. (2023) explored the use of graph-based encodings in volleyball.
  - Found that GNN-based models significantly outperformed traditional machine learning models.
  - Preserves the inter-player dependencies and sequential interaction
- Xenopoulos & Silva (2021) developed a general graph representation of game states that can be applied to a variety of sports.
  - Also found improved performance.
  - Able to answer "what if" questions that occur in sports through their modeling of player interactions.

Project Goal

Project Goal

Goal: Predict at-bat outcomes between pitchers and hitters.

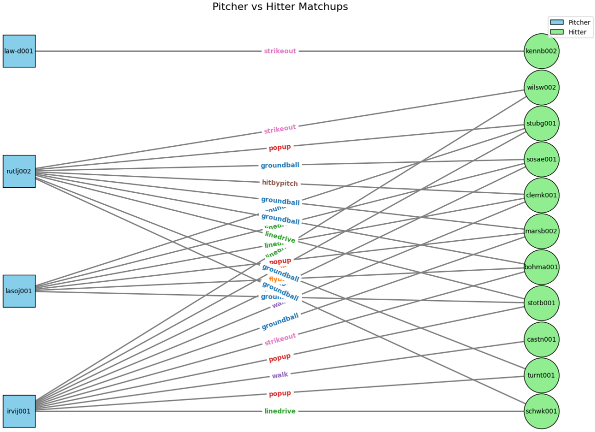
- Traditional baseball analytical methods rely on statistical models and machine learning approaches that often treat players in isolation or use aggregate statistics.
- GNNs have been applied to team-based sports such as basketball and volleyball, but has yet to be applied to the individual aspects of baseball.

Proposed Work: Model the pitcher-hitter matchup as a graph to leverage GNNs.

Data

Graph Structure

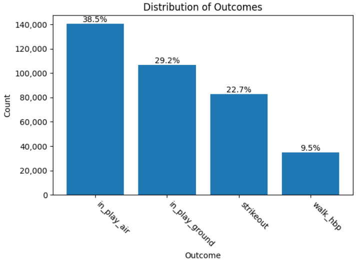
- Model the pitcher-hitter matchup as a bipartite graph.
- Nodes represent either a pitcher or a hitter.
- An edge exists between two nodes when a pitcher and hitter have previously faced each other.



Data Description

- Focuses on 2 seasons of data (2023-2024)
- Nodes: Player data collected from BaseballSavant
  - Pitcher Features (17): handedness, percent of each pitch type thrown, average fastball velocity, walk and strikeout rates.
  - Hitter Features (10): handedness, average exit velocity, average launch angle, swing percent, in-zone and out-zone contact percentage, whiff rate, strikeout and walk rates.
- Edges: Play-by-Play data collected from Retrosheet
  - Edge Features (5): inning, outs, indicator for each runner on base.
  - Predicted Outcome (4 classes): in_play_air, in_play_ground, strikeout, walk_hbp

Component	Count
Pitcher Nodes	1,555
Hitter Nodes	1,240
Edges	364,699



Model

Model Architecture

Selected GNN Architecture: Graph Attention Network (GAT)

- GAT enhances message passing between nodes by using attention mechanism.
- For each node, GAT computes attention coefficients with its neighbors based on their feature representations, allowing the model to focus on the most relevant connections.

Model Design:

- 2 layers of PyTorch's `GATv2Conv`
  - Applies to both edge directions.
  - Wraps in `HeteroConv` to keep the heterogeneous graph structure.
- Uses an attention-based edge classifier.

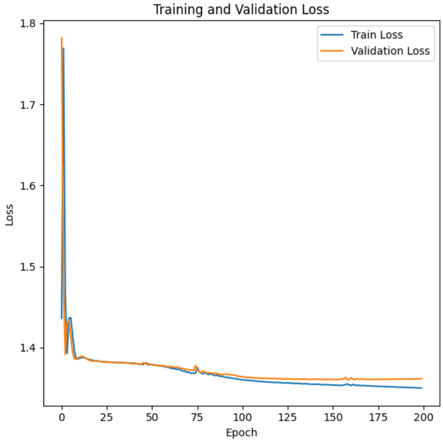
Training Process

- Split the data in train/validation/test sets.
  - Preserved the time-ordering of the edges when splitting.
- Class Balancing
  - Computes class weights using the "balanced" strategy to address class imbalance
  - Applies these weights to the CrossEntropyLoss function.
- Optimization Setup
  - Uses Adam optimizer with learning rate of 0.005
  - Tracks best model based on validation F1 score
- Training Loop
  - Runs for 200 epochs
  - For each epoch:
    - Forward pass through the model with training data
    - Computes loss using weighted CrossEntropyLoss
    - Performs backpropagation and parameter updates

Results

Training Results

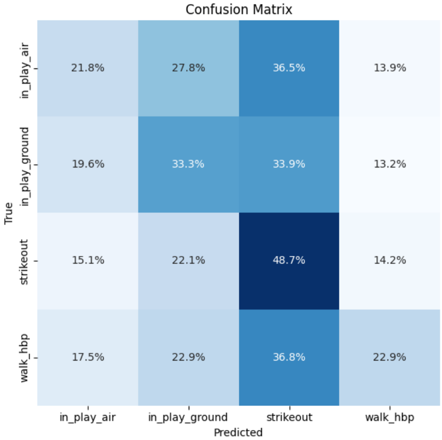
- The loss rapidly decreases and then flattens out around 1.35.
- The validation loss is similar to the training loss.
- Selected the model with the best macro-f1 score for the validation set which occured at epoch 157.



Results

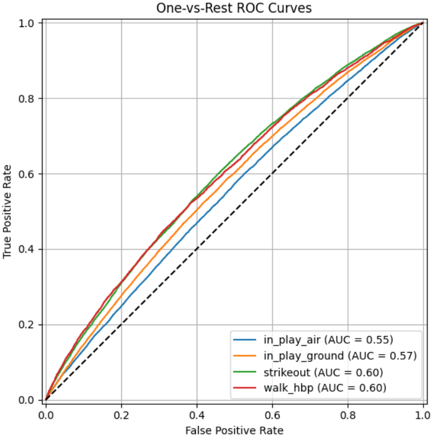
- The overall performance of the model in identifying the different classes is subpar.
- It is able to identify strikeouts the best and walks/hit-by-pitches the worst.

	precision	recall	f1-score	support
in_play_air	0.44	0.22	0.29	21105
in_play_ground	0.35	0.33	0.34	15878
strikeout	0.29	0.49	0.37	12705
walk_hbp	0.14	0.23	0.18	5018



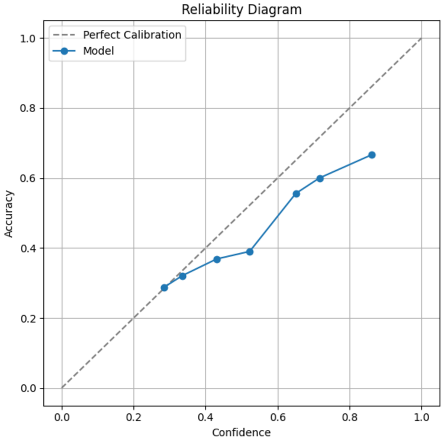
Results

- All predictions perform slightly better than random (AUC $= 0.5$).
- The strikeout and walk/hit-by-pitch predictions perform the best.
- The model struggles more with predicting balls in play.



Results

- The high brier score suggests there can be improvements in the calibration of the model.
- For confidence values below 0.4, the model appears well-calibrated (close to the diagonal)
- For confidence values above 0.4, the model shows overconfidence (the blue line falls below the diagonal)
    - At the highest confidence levels (0.8-0.9), the model predicts with about 67% accuracy despite being 90% confident



Component	Count
Brier Score	0.7305
ECE	0.0149

Future Work

Future Work

- The calibration of the model needs to improvement. The nature of the problem is difficult due to the overall distribution of outcomes for individual pitchers and hitter.
- Add temporal encoding to the model. The edges have a date attribute that has yet to be included in the model. This will potentially help identify trends to help with predictions.
- Another approach would be to include a LSTM framework to capture recent outcome distributions both for individual pitchers and hitters and pitcher-hitter pairs.

References

[1] K. Koseler and M. Stephan, "Machine Learning Applications in Baseball: A Systematic Literature Review," *Applied Artificial Intelligence*, vol. 31, no. 9–10, pp. 745–763, Nov. 2017. doi: 10.1080/08839514.2018.1442991.

[2] R. Tracy, H. Xia, A. Rasla, Y.-F. Wang, and A. Singh, "Graph Encoding and Neural Network Approaches for Volleyball Analytics: From Game Outcome to Individual Play Predictions," 2023, arXiv. doi: 10.48550/ARXIV.2308.11142.

[3] P. Xenopoulos and C. Silva, "Graph neural networks to predict sports outcomes," in *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA: IEEE, Dec. 2021, pp. 1757–1763. doi: 10.1109/BigData52589.2021.9671833.

[4] J. Silver, "Singlearity: Using A Neural Network to Predict the Outcome of Plate Appearances," *Baseball Prospectus*. Accessed: Mar. 10, 2025. [Online]. Available: https://www.baseballprospectus.com/news/article/59993/singlearity-using-a-neural-network-to-predict-the-outcome-of-plate-appearances/

[5] Retrosheet, "Game Sets by Season," *Retrosheet.org*. Accessed: Mar. 20, 2025. [Online]. Available: https://www.retrosheet.org/downloads/othercsvs.html

[6] MLB Advanced Media, "Custom Leaderboard: Pitchers (2015-2024)," *Baseball Savant*. Accessed: Mar. 20, 2025. [Online]. Available: https://baseballsavant.mlb.com/leaderboard

[7] MLB Advanced Media, "Custom Leaderboard: Batters (2015-2024)," *Baseball Savant*. Accessed: Mar. 20, 2025. [Online]. Available: https://baseballsavant.mlb.com/leaderboard