Progress Report

Selah Dean

Proposed Activities

The proposal listed the following activities to develop a graph neural network (GNN) model to predict at-bat outcomes in Major League Baseball (MLB) by leveraging a bipartite graph representation of pitcher-hitter interactions.

1. Data Collection and Preprocessing
    a. Determine timeframe to collect data from
    b. Create edge data set from Retrosheet play-by-play data
    c. Create pitcher and hitter node data sets from statistics on Baseball Savant
2. Model Selection
    a. Compare multiple GNN architecture such as GCN, GraphSAGE, and GAT
    b. Select one architecture to fully implement and tune
    c. Assess the importance of certain features
3. Evaluate Model
    a. Use cross-validation to assess accuracy
    b. Develop visualizations to help analyze the model's performance
    c. Assess calibration of prediction probabilities
4. Discussion
    a. Address the following questions
        i. Can a GNN effectively model pitcher-hitter interactions to improve at-bat outcome predictions?
        ii. What features contribute the most to the model?
        iii. How do historical matchups influence future matchups?
    b. Determine what further work can be done

Completed Activities

Data Collection and Preprocessing

I collected data from the 2015 through the 2024 MLB seasons. The edge dataset was constructed from Retrosheet's play-by-play, filtered to only retain rows that resulted in an end of the pitcher-hitter matchup (eliminates stolen bases, caught stealing, etc.). To create the outcome variable that the model will be predicting, I mapped event codes to a simplified set of eight classes: groundball, flyball, line drive, pop-up, bunt, strikeout, walk, and hit by pitch. Additional features for the edge attributes include inning (as well as top or bottom of the inning), the number of pitches in the at-bat, the number of times faced in the game, and if there is a runner on each base.

The pitcher and hitter data sets were created from data on Baseball Savant. For pitchers features include handedness, percent of each pitch type thrown, their average fastball velocity as well as their walk and strikeout rate. Pitchers without an average fastball velocity (i.e., position players who pitched in a game) were excluded. For the hitters features were handedness, average exit velocity, average launch angle, walk and strikeout rate. These datasets only include players who had at least ten plate appearances in the season.

Since player IDs differ between Retrosheet and Baseball Savant, I used the Chadwick Register to map players across sources. I converted the player ID from Baseball Savant to the player's Retrosheet ID so there can be mapping between the nodes and edges. Additionally, I dropped any rows in the edge data set that did not have corresponding

player in either the pitcher or hitter data set. When retrieving data from Baseball Savant the player must have at least ten plate appearances in the season, so these players had fewer plate appearances in the season.

After compiling the datasets, I standardized all continuous numeric values (grouped by year to adjust for seasonal variation) and encoded categorical features.

Incomplete Activities

Model Selection

I have tested several GNN models using data from a single season. So far, the most promising results have come from a Graph Attention Network. The next step is to extend the model to include data from multiple seasons and evaluate performance when incorporating temporal dynamics. I plan to have my final model complete by April 18th to give myself sufficient time to fully evaluate the results before my final presentation on April 23rd.

Evaluate Model

The model will be evaluated using standard multi-class classification metrics such as accuracy, F1-score, and confusion matrices. In addition, I will assess how well the predicted probabilities are calibrated using metrics such as expected calibration error and Brier score.

Discussion

This step will be completed after final model evaluation. I expect to address the core research questions and explore how player-specific and contextual factors influence matchup outcomes.

Challenges

Class Imbalance

This is a multi-class classification problem with substantial imbalance among the outcome classes. The first model I tested only predicted the two majority classes. To address this, I implemented class weighting in the loss to ensure that the model learns to predict minority classes more effectively.

Randomness of Outcomes and Evaluation Difficulty

There is a high degree of randomness in individual at-bat outcomes, and results between a given pitcher and hitter can vary over time. This makes strict accuracy-based evaluation difficult. To address this, I will place emphasis on evaluating the model's calibration and whether the predicted probabilities for each outcome match the observed frequencies. This approach allows me to better assess the quality of the model's probabilistic output in a setting where perfect predictions are inherently difficult.

Temporal Matching

Node features are season-based, but edge data is specific to individual dates. Because daily pitcher/hitter data is unavailable, all node features are matched by season. As a result, the model may miss short-term trends or streaks in player performance.

Changes in Scope

There is a slight change in the scope of the project to focus primarily on one GNN architecture instead of implementing various architectures. The original research questions and objectives remain unchanged.

References

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

[2] MLB Advanced Media, "Custom Leaderboard: Pitchers (2015-2024)," Baseball Savant. Accessed: Mar. 20, 2025. [Online]. Available: https://baseballsavant.mlb.com/leaderboard/custom?year=2024%2C2023%2C2022%2C2021%2C2020%2C2019%2C2018%2C2017%2C2016%2C2015&type=pitcher&filter=&min=10&selections=player_age%2Ck_percent%2Cbb_percent%2Cbatting_avg%2Cslg_percent%2Con_base_percent%2Cisolated_power%2Carm_angle%2Cn_ff_formatted%2Cn_sl_formatted%2Cn_ch_formatted%2Cn_cu_formatted%2Cn_si_formatted%2Cn_fc_formatted%2Cn_fs_formatted%2Cn_kn_formatted%2Cn_st_formatted%2Cn_sv_formatted%2Cn_fo_formatted%2Cn_sc_formatted%2Cn_fastball_formatted%2Cfastball_avg_speed&chart=false&x=player_age&y=player_age&r=no&chartType=beeswarm&sort=xwoba&sortDir=asc

[3] MLB Advanced Media, "Custom Leaderboard: Batters (2015-2024)," Baseball Savant. Accessed: Mar. 20, 2025. [Online]. Available: https://baseballsavant.mlb.com/leaderboard/custom?year=2024%2C2023%2C2022%2C2021%2C2020%2C2019%2C2018%2C2017%2C2016%2C2015&type=batter&filter=&min=10&selections=player_age%2Ck_percent%2Cbb_percent%2Cbatting_avg%2Cslg_percent%2Con_base_percent%2Cisolated_power%2Cexit_velocity_avg%2Claunch_angle_avg&chart=false&x=player_age&y=player_age&r=no&chartType=beeswarm&sort=xwoba&sortDir=desc