Classifying 2024 MLB Pitch Types

Alexander J. Sutherland

Overview: No two pitchers truly throw the same pitch the same way. For example, Tim Hill threw his four-seam fastball with an average of 17.9in of arm-side movement, 4.4in of vertical movement, and at 90.7 MPH. Mason Miller, on the other hand, threw his four-seam fastball with 9.7in of arm-side movement, 16.6in of vertical movement, and at 100.9 MPH. In this project, we classify pitch types using observable data, paying special attention to the inclusion/exclusion of pitcher IDs and the differences between classical machine learning models and a simple neural network.

Project Stakeholders: MLB, MiLB, Professional baseball organizations, Professional baseball players, Baseball fans.

Data: We use pitch data from the entire 2024 MLB season, which is originally provided and hosted by *Baseball Savant* (MLB Advanced Media, LP) and accessed through *pybaseball* (LeDoux & Schorr).

Methods: For our classical models, we considered k-nearest neighbors, decision trees, and several ensemble methods with 10-fold cross validation and hyperparameter tuning via grid search. These models were compared against each other and a baseline model, where the odds of a pitch being predicted as a given class where given by the overall distribution of the test dataset.

For the neural network, we constructed a basic neural network with multiple linear layers using pytorch.

Key Performance Indicators (KPIs): For all our models, we prioritize model accuracy. We additionally track precision, recall, and F1-score as secondary metrics.

Conclusions and Future Work: When using a pitcher identifier, almost all classical models performed very well - four of the five (non-baseline) models had an accuracy over 95%. Without the pitcher identifier, our one classical model still performed well (with an accuracy of 84%), but with a noticeable drop in accuracy. The specific neural network in question did not perform as well (with an accuracy of 67%), but there is reason to believe that a more sophisticated architecture could lead to better results.

References

LeDoux, J., & Schorr, M. (2024, Oct 03). *Pybaseball Github Repository Readme*. Retrieved from Pybaseball Github Repository: https://github.com/jldbc/pybaseball

MLB Advanced Media, LP. (2024, Oct 03). *Statcast Search*. Retrieved from Baseball Savant: https://baseballsavant.mlb.com/statcast_search