Predicting College Baseball Pitcher Performance Using Machine Learning

Portfolio Project 2

By Mark Fleet

DSC-680, Professor Catherine Williams

Bellevue University

Business Problem:

The project aims to predict the performance of college baseball pitchers based on historical data from 2000 to the most recent year. The analysis can assist college baseball teams, coaches, and talent scouts in identifying key factors contributing to a successful pitcher and making informed decisions on player recruitment and development.

Datasets:

Historical Pitcher Data: This dataset will include historical data of college baseball pitchers, including individual statistics such as earned run average (ERA), strikeouts, saves, innings pitched, and other relevant metrics. Data sources include the NCAA website (https://www.ncaa.com/stats/baseball/d1) and other relevant baseball statistics websites.

Methods:

To complete this project, the following methods will be employed:

- Data Collection: Collect and preprocess the historical pitcher data for Division 1 college baseball from 2000 to the most recent year.
- Data Cleaning and Preprocessing: Clean the data, handle missing or incomplete data, and convert categorical variables into numerical variables if needed.
- Feature Engineering: Create new features that may be relevant for predicting pitcher performance, such as the pitcher's win percentage, strikeout-to-walk ratio, or WHIP (Walks plus Hits per Inning Pitched).

• Machine Learning Model: Build various machine learning models, such as linear regression, decision trees, or random forests, and evaluate their performance using appropriate metrics such as mean squared error, R-squared, mean absolute error.

• Model Interpretation and Presentation: Analyze the importance of different features in the selected model and present the findings and implications for college baseball teams and coaches.

Ethical Considerations:

Potential ethical considerations for this project include:

• Ensuring fairness and avoiding bias in the analysis and model predictions.

• Respecting the privacy of the players and avoiding the use of any sensitive personal information.

• Transparently disclosing the limitations and uncertainties of the predictions.

Challenges/Issues:

• Data Availability: Ensuring that accurate and up-to-date historical data is accessible for all years.

• Data Consistency: Ensuring the data is consistent across all years, as different sources might have different formats and column names.

• Model Complexity: Managing the complexity of machine learning models and avoiding overfitting.

• Ethical Concerns: Addressing potential biases and ethical concerns associated with player performance predictions.

References:

• NCAA. (n.d.). Baseball Statistics and Rankings. Retrieved from https://www.ncaa.com/stats/baseball/d1

• Baseball-Reference. (n.d.). College Baseball Stats on Baseball-Reference.com. Retrieved from https://www.baseball-reference.com/schools/

• D1Baseball. (n.d.). College Baseball Rankings, Scouting Reports, and Analysis. Retrieved from https://d1baseball.com/

• (More resources to come as EDA continues)