Final Project Proposal
INFO 523
Dr. Greg Chism
Austin Cortopassi, David Pelley, Nathan Harville

High-Level Overview

We will build a machine learning model to predict whether a Major League Baseball player will be inducted into the Hall of Fame based on their career statistics and achievements.

Project Goals

Our goal is to understand what motivates Hall of Fame inductions in Major League Baseball. By utilizing the Lahman Baseball Database, we can develop a prediction model using historical data for baseball players dating back to the 1800s. We aim to realize patterns that may influence Hall of Fame voters.

We chose to use the Lahman Baseball Database because of the level of detail provided on all facets of a player's game.

Below are several datasets provided by the Lahman Baseball Database:

- **People**: Contains biographical information for each player (e.g., name, position, debut/final year).
- **Batting**: Season-level batting statistics (e.g., HR, AVG, RBI).
- **Pitching**: Season-level pitching data (e.g., ERA, strikeouts, wins).
- **Fielding**: Season-level defensive stats by position (e.g., errors, assists, putouts).
- **Teams**: Team-level seasonal statistics (e.g., win/loss records).
- **BattingPost**, **PitchingPost**, **FieldingPost**: Postseason statistics, used to measure clutch or playoff performance.
- **HallOfFame**: Voting results including year, ballots, votes, and whether the player was inducted.

Each table contains a **playerID** column that allows us to join across sources. The data spans from the 19th century through the most recent MLB season, offering rich historical coverage.

Weekly Planning:

Week 1: Data Preparation and Exploration - Pre-Processing

Loading datasets, generating/cleaning DataFrames, and exploratory data analysis (EDA).

Austin: Batting, BattingPost, and People. Merge all 9 datasets when done.
David: Pitching, PitchingPost, and HallofFame. Cleaning and Prep
Nate: DataFrame - Fielding, FieldingPost, Team. Cleaning and Prep

Week 2: Feature Engineering

Aggregate batting, pitching, and fielding to career-level stats; join with labels.

Austin: Aggregate batting and battingPost statistics to career totals. Engineer derived features such as AVG, OPS, and HR/year.

David: Aggregate pitching and pitchingPost data. Create key metrics like WHIP, quality starts, strikeouts/walks ratio (K/BB).

Nate: Aggregate fielding and fieldingPost data. Engineer features such as games played per position and WAR.

Week 3: Train Models and Tune Parameters

Train baseline models using logistic regression and decision trees. Tune parameters and evaluate output.

Austin: Work with David to implement logistic regression, decision tree models.

David: Work with Austin to implement logistic regression, decision tree models.

Nate: Implement and evaluate tuning methods for logistic regression and decision tree models.

Week 4: Final Model Evaluation and Reporting

Evaluate final model output as well as data visualizations. Write a final report and make our presentation.

Austin: Create visualizations of model results (e.g., feature importances, ROC curve). Summarize EDA findings visually for the presentation.

David: Write the technical summary of the model development process. Document model evaluation and limitations.

Nate: Write conclusions and discussion on what features were most predictive. Format the report and contribute to building presentation slides.

Project Repository Structure:

data/ Raw and processed data.

notebooks/ Notebooks for batting, pitching and fielding.

src/ Feature and model integration.

figures/ Data visualizations.

reports/ Final report and presentation.