Hierarchical Ensemble Model for Cy Young Prediction

Charles Benfer

April 22, 2025

Outline

1. Motivation
2. Data Collection
3. Level 1: Pitch-Level Modeling
4. Level 2: Game-Level Sequence Modeling
5. Level 3: Bayesian Hierarchical Modeling
6. Meta-Model Stacking
7. Training Results
8. Current Season Prediction
9. Conclusion & Future Work

Cy Young Award

- The Cy Young award is given out each year to the best pitcher in the American League and the National League of Major League baseball
- The award is voted on by the BWAA, Baseball Writers Association of America
- Correctly predicting the Cy Young winner can lead to serious financial gains.

Data Scraping

- Leveraged Pybaseball package and TJ Nestico's MLB Data Scraper to obtain pitch level and season total data for all pitchers to finish in the top 10 of Cy Young voting in both leagues
- We utilize data from the season prior to a pitcher's Cy young placement for prior information, alongside data from the season they placed in the top 10. This is in an attempt to capture previous trends in the relevant models, as well as trends in the current season.

Level 1: Pitch-Level Ensemble

- **Swing Classifier**: Predict $P(\mathrm{swing})$ using RF.
- **Whiff Classifier**: Predict $P(\mathrm{whiff} \mid \mathrm{swing})$.
- **Exit Velocity Regressor**: Predict $E(\mathrm{exit\_vel} \mid \mathrm{swing} \,\&\, \neg\mathrm{whiff})$.
- **Pitch Score**: $P_{\mathsf{swing}} \times (1 - P_{\mathsf{whiff}}) \times \mathrm{exit\_vel}$.
- *Hyperparameters*: Tuned via RandomizedSearchCV.

Level 2: LSTM on Game Sequences

- Aggregate pitch scores to per-game mean: $\mathrm{pitch\_score\_mean}$.
- Form sequences of length 5 (past 5 games).
- **LSTM**: Input shape $(5, 1)$, output next-game RBI prediction.
- *Hyperparameters*: Units, dropout, learning rate, batch size, epochs tuned via Keras Tuner.

Level 3: Bayesian Hierarchy

- Model game-level RBI predictions per pitcher-season:

$$a_i \sim N(\mu, \sigma), \quad y_{ij} \sim N(a_i, \epsilon).$$

- **Outputs**: Posterior means a_i as pitcher effects.
- *Inference*: PyMC & NUTS sampling.

Meta-Model: Logistic Stacking

- Features per season:
  - $L_1 =$ season average pitch_score.
  - $L_2 =$ season average predicted RBI.
  - $L_3 =$ Bayesian effect a_i.
- **Logistic Regression** on $[L_1, L_2, L_3]$.
- *Hyperparameter C* tuned via GridSearchCV.

Level 1 Metrics

Model	Accuracy	AUC	MSE	R^2
Swing Classifier	0.7956	0.8746	–	–
Whiff Classifier	0.7607	0.7398	–	–
Exit Velocity	–	–	196.4619	0.0917

Level 2 & Meta Metrics

Model	Accuracy	AUC	MSE
LSTM	–	–	.0094
Meta Model	0.929	0.618	–

Predicting Current Cy Young

1. Scrape & score pitch-level.
2. Aggregate to game-level, predict RBIs via LSTM.
3. Map Bayesian effects (or use global mean fallback).
4. Form meta-features and predict probabilities.

So Who's Gonna Win 2025?

Top 5 Predicted Cy Young Winners for 2025 are:

- Hunter Greene - Cincinnati Reds (7.74%)
- Spencer Schwellenbach - Atlanta Braves (6.72%)
- Framber Valdez - Houston Astros (4.20%)
- Cole Ragans - Kansas City Royals (3.68 %)
- Luis Severino - Athletics (3.68%)

Conclusion

- Modular, interpretable multi-level ensemble.
- Incorporates raw pitch data, time series, and hierarchical effects.
- *Future*: Add win probability, lineup context, real-time updates.