Chicago Cubs – Baseball Analytics

Patrick Isom, Ram Muthurangan, Anvith Chigurupati, Jobe Holderread, Grant Johnson, Colin Robinson, Ananya Bhumireddy, Atul Narahari, Calester Rem, Sairah Ghule, Sujay Kodam, Michael Wolf

PURDUE UNIVERSITY | The Data Mine

Purpose & Background

Managing player incentives is crucial for the Cubs' front office to stay under the luxury tax threshold.

- In 2023, Cody Bellinger's Comeback Player of the Year win triggered a $1M bonus, highlighting the need for better financial planning.
- This project builds a machine learning model to predict award winners, using player stats and historical trends
- Accurate predictions help optimize payroll decisions and maintain a competitive roster
- By leveraging data-driven insights, the Cubs can make smarter contract offers and allocate resources.

Data

Data Processing
- Extracted data from SQL and converted it to CSV files.
- Imported batting, fielding, and pitching data for analysis.
- Encoded award names as binary values for logistic regression.
- Normalized batting stats by dividing by plate appearances.

Key Attributes
- Primary Key: player_id (Unique identifier for each player)

Main Variables Used:
- WAR (Wins Above Replacement) – Measures overall player value.
- Batting Average per Inning – Evaluates a hitter's efficiency.
- ERA (Earned Runs per Inning) – Assesses a pitcher's performance.

Target Variable
- Votes: Number of votes received for an award.
- Award_name: Binary to indicate whether a player won an award.

Results

Logistic Regression

Predicted 2023 Cy-Young Winners

AL	NL
Felix Bautista	Blake Snell
Chris Martin	Clayton Kershaw
Kyle Bradish	Justin Steele

XGBoost

Predicted 2023 Cy-Young Winners

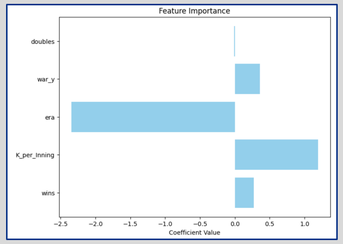
AL	NL
Gerrit Cole	Blake Snell
Pablo López	Spencer Strider
Kevin Gausman	David Bednar

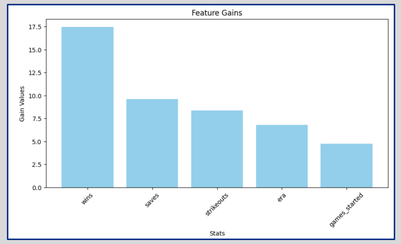
Comparison

Actual 2023 Cy-Young Winners

AL	NL
Gerrit Cole	Blake Snell
Sonny Gray	Logan Webb
Kevin Gausman	Zac Gallen

Features



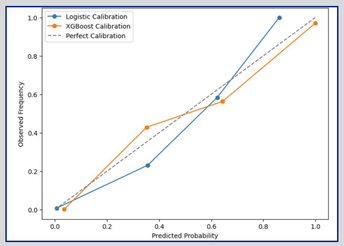


Importance Values
- Log. Reg: coefficient associated, how much that feature implies
- XGBoost: model gain per feature

Feature Selection
- Log. Reg.: hand-picked features, not all useful
- XGBoost: chosen by algorithm, selects most important by gain

Comparison
- Some similar features (wins, era)
- More value lent to wins and strikeouts in XGBoost

Comparison



Calibration Plot
- Compares observed values of probabilities versus perfect frequency
- Used to see effectiveness of model

Analysis
- Log. Reg. predicts low for lower chances of winning an award
- XGBoost predicts lower for higher chances
- Both are close to perfect line, XGBoost moves closer to perfect as time goes on
- Overall, XGBoost is better predictor

Overall Results
- Log. Reg.: 1/6 on 2023 Cy Young Winners
- XGBoost: 3/6 on 2023 Cy Young Winners

Which to choose?
- Based off results, XGBoost predicts better
- XGBoost is faster

Logistic Regression

Purpose
- Help users predict through binary classification.

Results
- Results are very understandable for stakeholders to understand the most significant attributes.

Use Cases
- These models perform very well on small to medium datasets with well-defined categories of data.
- The predictive power of the logistic regression can help users analyze historical data on player performance and model statistical probabilities.

XGBoost

Decision Trees
- Builds and tests many decision trees
  - Finds features in data to split on
  - Splits on that feature, retests to find new split
  - Each tree is tested and an improved tree is created based on the results (Gradient Boosting)

Uses
- Binary (ex. Win/Loss)
- Classification (top 10, top 5, etc.)
- Suited for large datasets
- Black box model

Comparison

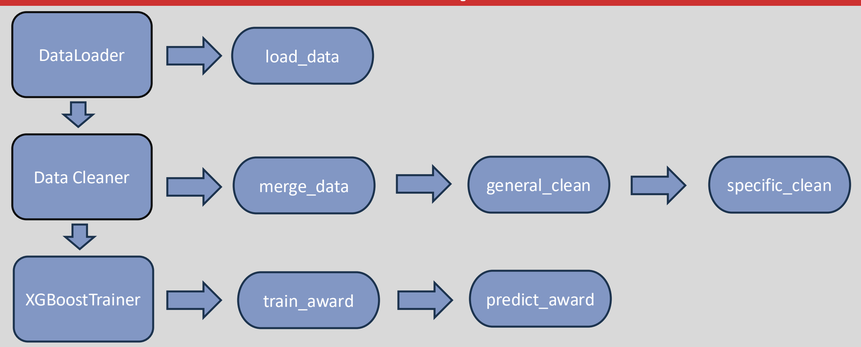
Switched from logistic regression to XGBoost and saw the Following improvements:

Enhanced Accuracy
- Gradient boosting refines predictions by combining multiple weak models.

Efficiency
- Better handling of large datasets, making it suitable for more complex tasks.

XGBoost Pipeline

BEFORE
- Running individual cells of code
- Non-uniform data cleaning between sub-teams
- No standardized training/testing of models
- Directories being cluttered and spread out



AFTER
- Grouped code functions for easy cleaning and model creation
- Standardized cleaning for all data
- One function to run for training and testing models
- Functions for each team located in one file

Future Goals

Interface for Cubs Model:
- Create customizations for different models and better usability

Extra Features for Model:
- Experiment with clutch performance metric, injury history impact, and etc

Improve Player Predictiveness:
- Complement XGBoost and capture more complex patterns in player performance

Summary

Logistic Regression & XGBoost
- Integrated XGBoost models over original logistic regressions
- Significantly improved model efficiency and accuracy

Pipeline
- Optimized the pipeline by streamlining code and standardizing processes.

Proof of Concept
- Learned what awards can be modeled for the Cubs to improve upon our work

Acknowledgements

Jeremy Frank & The Chicago Cubs
The Data Mine Staff:
Jessica Gerlach, Ashley Arroyo