Using Decision Trees for Baseball Run Prediction

Final Presentation

Lance Kevin, Ian McDowell, Michael Templeton

Problem Statement

Problem: How can South Carolina Baseball improve their team's batting discipline to have a higher chance at winning games?

Plan: Our project will generate a *hit-ability score* for each pitch.

Data: College Trackman dataset provided to us by the team

Expected Outcome: model that determines plate discipline for players which will allow for predictions of run-value for any given pitch, which will provide insights into a player's plate discipline.

A decision tree with insightful decision nodes on what aspects of a pitch to focus on when deciding whether or not to swing.

Challenges

- Inexperience with machine learning
- What hyperparameters should we test
- Learning to program with SciKit Learn Python Library
- Fitting our data into a model

Related Works

- Tango Lichtman Dolphin blog article on swing/take, and run expectancy matrix
- QOP^{tm} (quality of pitch) and the Griner Index.

Data

- Pitching Dataset provided by UofSC Baseball Team
- Preprocessed
- 1,029,479 useable data points
  - 720,635 (70%) used for training
  - 308,844 (30%) used for testing

Methods

- Our approach uses decision trees and random forests.
  - The goal of these models is high explainability so that we can use our model to determine important factors of hit-ability
- We also trained a neural network (Multi-Layer Perceptron)
  - The goal of this model is to examine the explainability vs accuracy tradeoff

Experiments: Decision Trees

- Total models trained: 1400
- Test Accuracies:
  - Best model - 63.60%
  - Average model - 57.84%

- Color Meanings:
  - Orange: Ball Called
  - Pink: Correct Swing
  - Blue: Strike Swinging/ Strike Called

- Important Factors:
  - Plate Location Height
  - Plate Location Side
  - Vertical Break
  - Ball Strike Number

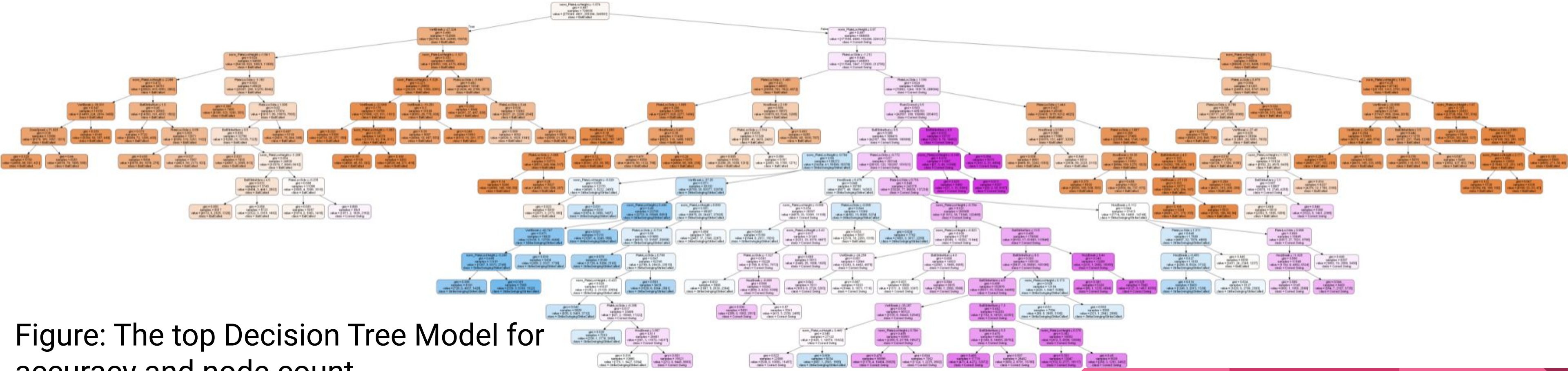

Figure: The top Decision Tree Model for accuracy and node count

Experiments: Random Forest

- Total models trained: 144
- Test Accuracies:
  - Best model - 65.95%
  - Average model - 63.66%

- On average gave more complex trees than the decision tree model

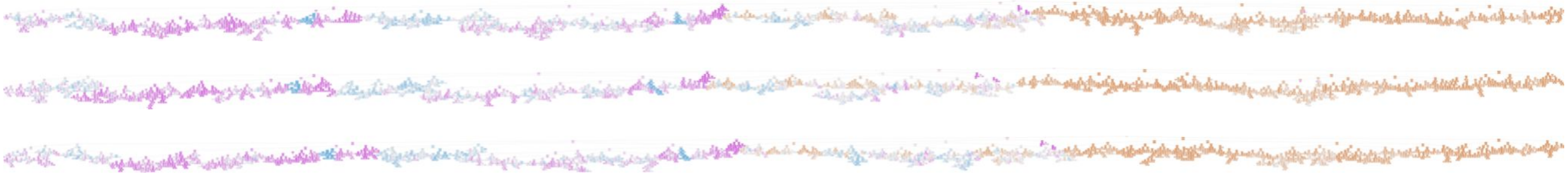

Figure: 3 decision trees from one of the higher accuracy forests

Experiments: Random Forest, Pitch-Separated

- Total models trained: 24 (6 per pitch type)
- Test Accuracies:
  - Fastball
    - Best model - 66.19%
    - Average model - 65.33%
  - Curveball
    - Best model - 65.60%
    - Average model - 65.36%
  - Change Up
    - Best model - 63.10%
    - Average model - 62.77%
  - Slider
    - Best model - 63.43%
    - Average model - 63.18%

- Insights:
  - Fastball is the most accurate, likely due to having the majority (~56%) of samples

Experiments: Neural Network

- Total models trained: 42
- Test Accuracies:
  - Best model - 68.28%
  - Average model - 67.76%

- Not as accurate as we hoped (>70%) but still gained insights
  - Average accuracy higher and deviates less from average than decision tree and random forest models
  - There is a clear tradeoff between explainability and accuracy

Broader Impact

1. South Carolina Baseball team will have a useful way to improve player performance and hopefully lead to more wins
2. South Carolina Measure players plate discipline and make training and coaching adjustments to improve player and team performance
3. The data we have and the inexperience of the group with machine learning, lower accuracy of 68%
4. Future work Train Model for different strikes zones, left hand to right hand matchups, defensive shifts and applying method in real life.

References

1. Tango Lichtman Dolphin.. (2019 September 23). *Statcast Lab: Swing/Take and a Primer on Run Value.* Tantotiger. http://tangotiger.com/index.php/site/article/statcast-lab-swing-take-and-a-primer-on-run-value
2. *Quality of Pitch.* https://en.wikipedia.org/wiki/Quality_of_Pitch

3. Jason Wilson & Jarvis Greiner (2014) A Curveball Index: *Quantification of Breaking Balls for Pitchers*, CHANCE, 27:3, 34-40, DOI: 10.1080/09332480.2014.965629 https://www.tandfonline.com/doi/full/10.1080/09332480.2014.965629?scroll=top&needAccess=true