Model Card: Predicting MLB Hit Distance

Ian Curtis

2022-12-08

Model Details

- Developed by Ian Curtis for a university class project (Fall 2022)
  - CIS 635: Knowledge Discovery and Data Mining
- Consisted of three models (two of which are included in this card)
  - Decision Tree classification (factor as independent variable)
    - Complexity parameter of 0.00001
    - Max depth of 6
  - Neural Network (factor as independent variable)
    - Hidden layer size of 5
    - Decay of 1.0e-7
- Trained on MLB 2022 regular season data

Intended Use

- These models were intended for researchers exploring different models to predict hit distance or other various metrics
- These models are not intended for serious uses or for predicting hit distance. The purpose of these models were to determine how they would perform not for optimization or finding the "best" model.
- These models are not suited for sports other than baseball; may also not generalize to other MLB seasons or postseasons

Factors

- The performance of these models may vary in other sports, different seasons, postseasons, or with other variables. These models may also not perform as well on minor league baseball or other non-major-league areas.
- Evaluation factors are launch speed (speed of a ball off of a bat) and launch angle (the angle of a ball off of a bat). Other factors were available in the dataset and were tested but were not seen to be beneficial for model accuracy. Thus the model that contained only launch speed and launch angle was chosen (one of a decision tree and one of a neural network).

Metrics

- The sole metric used to evaluate model accuracy was total accuracy which is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

where T = True, F = False, P = Postitive, and N = Negative. Accuracies were compared for train and test data.

Training and Evaluation Data

- The overall data was split in the preprocessing step so that each model used the same train and test data. 85% of the observations were in the training data (`train.csv`) and the remaining 15% became the test data (`test.csv`).

Ethical Considerations

- No ethical considerations need to be made. Data is sourced publicly.
- Again, the dataset and models here are not intended to be used in a serious, predictive sense as the project's goal was to determine a potentially effective model. The results from these exact models may not provide direct useful insight to a baseball team.

Caveats and Recommendations

- Further testing is needed to determine the extent to which these models perform the best (or how other models outperform them).
- Future work can expand the predictor variables, include data from prior seasons, and try to predict hit distance using only data that can be obtained before a batter hits a ball.

Quantitative Analyses

- The training accuracy for the chosen model using the decision tree algorithm was 0.7391 and the corresponding test accuracy was 0.7322.
- The training accuracy for the chosen model using the neural network algorithm was 0.8124 and the corresponding test accuracy was 0.8114.
- The following graphic shows train and test accuracies for all models testing using the decision tree and neural network algorithms.
  - Full model uses all variables
  - Reduced 1 uses all variables except launch speed and launch angle
  - Reduced 2 (the subject of this card) uses only launch speed and launch angle

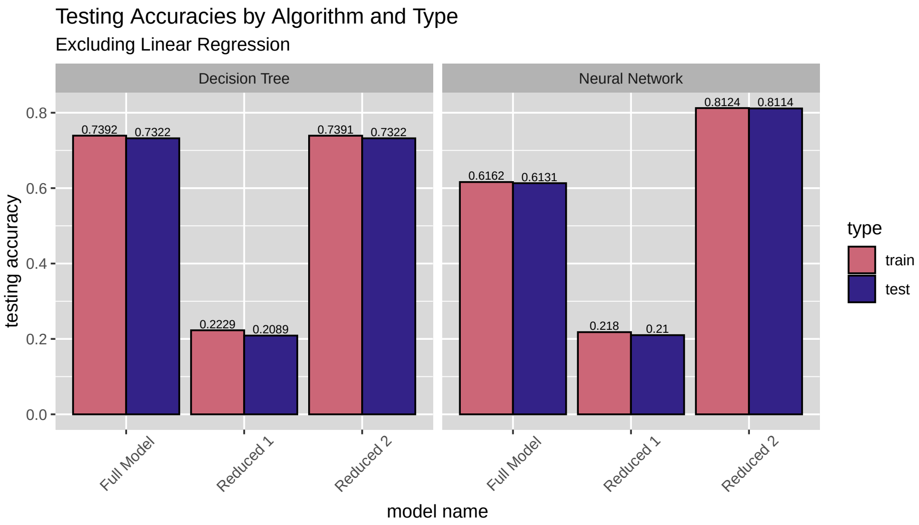

Figure 1: Residual Plot for Reduced 5