PREDICTING PITCHES IN THE MLB

CHS 788 - FINAL PROJECT

ALEXIA SPOON AND WILL BLISS

GOAL

We would like to predict a pitch type before it is thrown

We will be focusing on 2015 Astros' pitcher Dallas Keuchel

Data source: https://www.kaggle.com/pschale/mlb-pitch-data-20152018

ABOUT DALLAS KEUCHEL (2015)

- Starting pitcher for the Houston Astros
- Awards in 2015:
  - 3x AL Pitcher of the Month
  - AL All-Star Selection and Starter
  - AL Cy Young Winner
- 2.48 ERA
- 20 – 8 Record
  - 15 – 0 at Home

THE DATA – FEATURES (11 TOTAL)

Ball count:	Strike count:	Number of outs:	Pitch count per batter:	Runner on first:	Runner on second:	Runner on third:	Inning number:	Batter's stance:	Bottom/Top of inning:	Run differential:
• Categorical data • Made up of 0, 1, 2, and 3	• Categorical data • Made up of 0, 1, and 2	• Categorical data • Made up of 0, 1, and 2	• Categorical data • Made up of 1, 2, ..., 10	• Binary data • Made up of 0 (false) and 1 (true)	• Binary data • Made up of 0 (false) and 1 (true)	• Binary data • Made up of 0 (false) and 1 (true)	• Categorical data • Made up of 1, 2, ..., 9	• Binary data • Made up of 0 (left) and 1 (right)	• Binary data • Made up of 0 (bottom) and 1 (top)	• Categorical data • Ranges from -8 to 15*

* We combined the original dataset's features **pitcher's team score** and **batter's team score** to make the ***run differential*** feature, calculated as **pitcher's score – batter's score**

THE DATA - OUTCOME

- Pitch Type (Binary) – Fastball (F) and Off-Speed (O)
  - Pitches classified as Fastball: 4-Seam Fastball, 2–Seam Fastball, Cutter
  - Pitches classified as Off-Speed: Changeup, Slider

METHODS PERFORMED

01

Linear Discriminant Analysis

02

Classification Tree

03

Conditional Inference Tree

04

Random Forest

05

ADA Boost

LINEAR DISCRIMINANT ANALYSIS

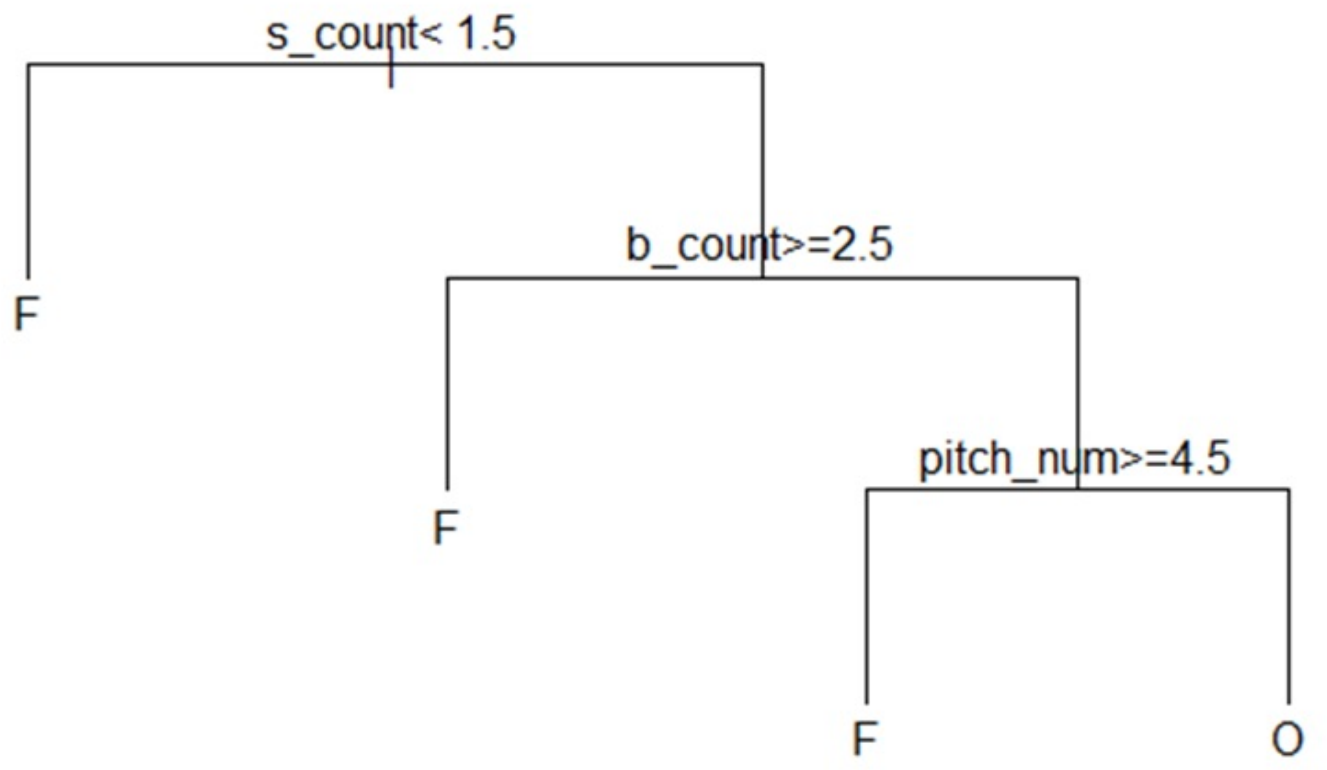
Packages used:	• `MASS`
Functions used:	• lda()
Cross validation:	• 10-fold
Estimated test error:	• 32.50%

CLASSIFICATION TREE

Packages used:	• `rpart`
Functions used:	• rpart()
Cross validation:	• 10-fold
Estimated test error:	• 31.95%

CLASSIFICATION TREE

- s_count:
  - Number of strikes in at-bat
- b_count:
  - Number of balls in at-bat
- pitch_num:
  - Number of pitches in at-bat
- F:
  - Outcome of fastball
- O:
  - Outcome of off-speed

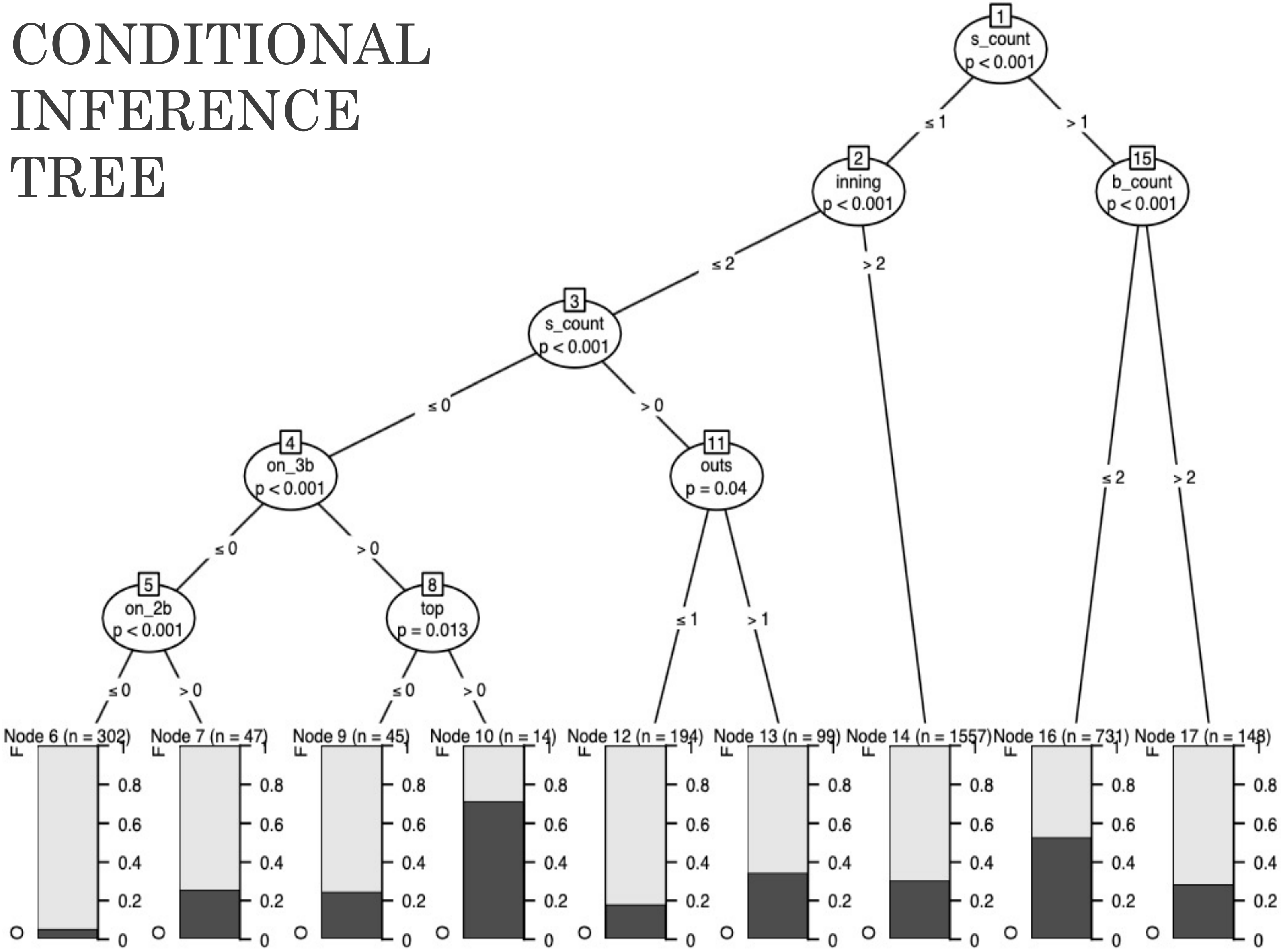


CONDITIONAL INFERENCE TREE

Packages used:	• `partykit`
Functions used:	• ctree()
Cross validation:	• 10-fold
Estimated test error:	• 31.38%

- s_count:
  - Number of strikes in at-bat
- b_count:
  - Number of balls in at-bat
- inning:
  - Inning number in at-bat
- outs:
  - Number of outs in at-bat
- on_3b:
  - 1 if runner is on 3rd, 0 if not
- on_2b:
  - 1 if runner is on 2nd, 0 if not
- top:
  - 1 if top of the inning, 0 if not
- F:
  - Outcome of fastball
- O:
  - Outcome of off-speed

CONDITIONAL INFERENCE TREE



RANDOM FOREST

Packages used:

- `randomForest`

Functions used:

- randomForest()

Attempted tuning parameters:

- mtry = 1, 2, ..., 11
- ntree = 500, 1000

Best tuning parameters:

- mtry = 2
- ntree = 1000

Estimated test error:

- 32.64%

ADA BOOST

Packages used:	• `gbm`
Functions used:	• gbm() • gbm.perf() • predict.gbm()
gbm() tuning parameters:	• distribution = “adaboost” • n.trees = 20000
predict.gbm() tuning parameters:	• n.trees() = 101
Estimated test error:	• 31.50%

CONCLUSION

- Best method: Condition Inference Tree (31.38% estimated test error)
  - We believe that this method performed the best because it has a realistic approach to predicting the pitch type. We also believe that the Conditional Inference Tree performed better than the Classification Tree because the Cond. Inf. Tree utilizes p-values when making decisions.
- Worst method: Random Forest (32.64% estimated test error)
  - While the Random Forest was the worst method, it was only out-performed by the Cond. Inf. Tree by 1.26%. We do not have any reasoning as to why this method performed the “worst”.
- Computational Challenges:
  - It should be noted that the results will vary when choosing a seed for randomizing training observations. For a different seed, we may get results that show different models perform better or worse.
  - We initially had issues formatting the data in ways that allowed the methods to be performed. For example, we had to narrow the outcome to Fastball and Off-speed as there were too many pitch-type outcomes in the beginning.
  - All methods ran about the same speed computationally.
  - R was a great computational software for this project. It was very user-friendly, and the libraries allowed us to perform all tests that we wanted to.