Bayesian Classification for NFL Defensive Coverages

BIG DATA BOWL

Matt Manner, Connor Nickol, Josh Gen

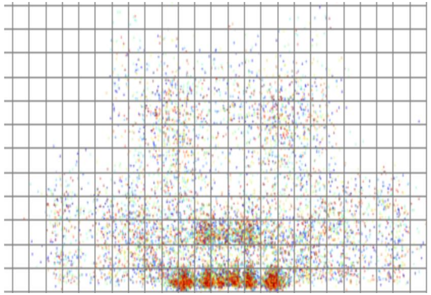
University of Virginia | School of Data Science

Background

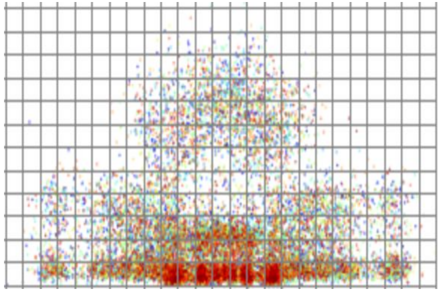
- NFL's yearly Big Data Bowl analytics competition
  - NFL releases Amazon Web Services (AWS) Next Gen Stats player tracking data
  - Our data: first 8 weeks of 2021 season
- Included data sets:
  - Players: name, unique ID, position for 1679 players
  - Plays: unique play ID, offensive/defensive team, pass coverage for 8557 plays
  - PFF: player, play, position for 188254 player-plays
  - Games: 10x/sec player & football location tracking data for each play
- Goal: predict defensive coverages from defensive player alignment pre-snap
  - Motivation: thousands of man-hours spent analyzing film for multi-billion dollar industry
  - Want appropriate uncertainty measures for this classification problem

Data Processing

- Defensive players only
- Only consider player locations at moment ball is snapped
- Use football location data to code player locations
  - Want player locations relative to the ball, rather than absolute position on field
  - Ball is always (0,0) for each play
- Transform location data from coordinates to numbered "box" system covering the useful playing field (sideline to sideline, line of scrimmage to 26 yards deep)



COVER 2 (n = 944)



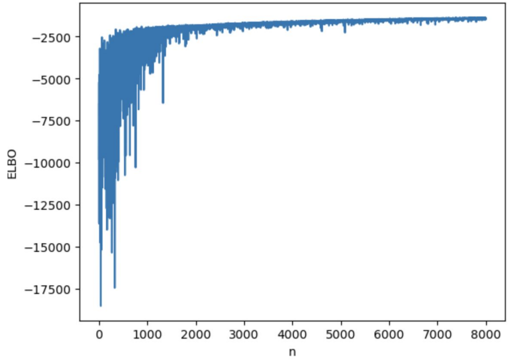
COVER 1 (n = 1759)

Each box used as a predictor variable, where value = number of players in box before snap

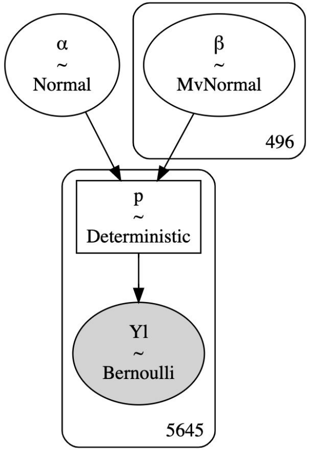
UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

METHODOLOGY

- Approx. 7800 observations (plays) and 496 predictors (field grids)
  - Sparse matrix: only 11 defensive players allowed per play so most field grids are empty
- Response variable is coverage type (7 levels)
  - One-hot encoded to enable 7 one vs. rest Bayesian logistic regressions
- Posterior for logistic regression coefficients obtained via Variational Inference
  - Computationally intractable, large data set



COVER 5 regression ELBO plot



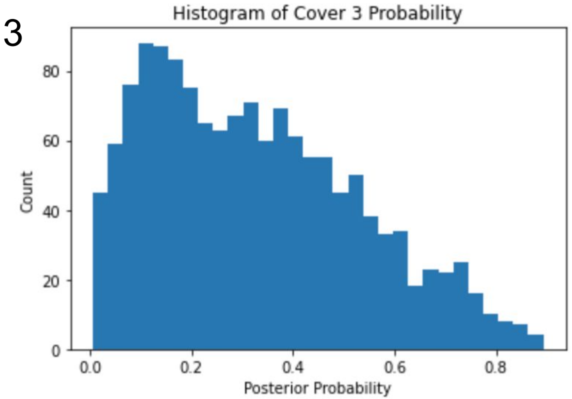
UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

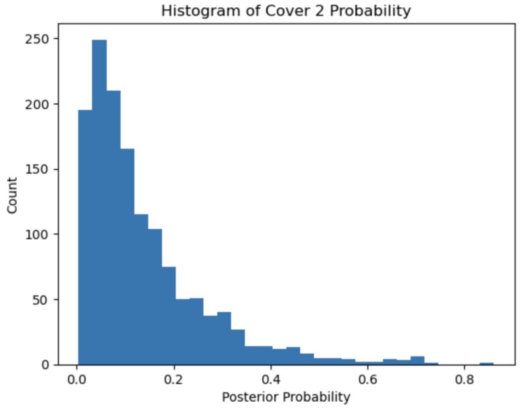
Code Reference for posterior prediction: https://juanitorduz.github.io/sklearn_pymc_classifier/

Results/Output

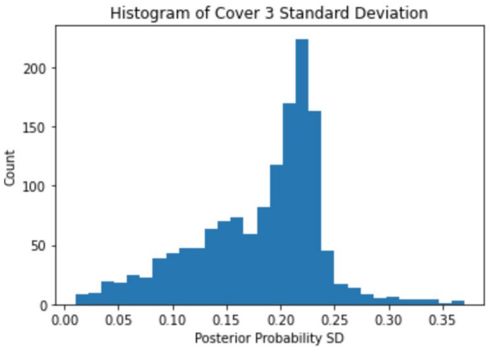
- Calculated 94% credible interval for each coverage probability for each sampled play
- Gauged accuracy by choosing Bayes optimal coverage (maximum probability)
  - Accuracy: **47.03%** vs. 14.29% from random guessing (1/7) vs. 33.2% from always guessing Cover 3

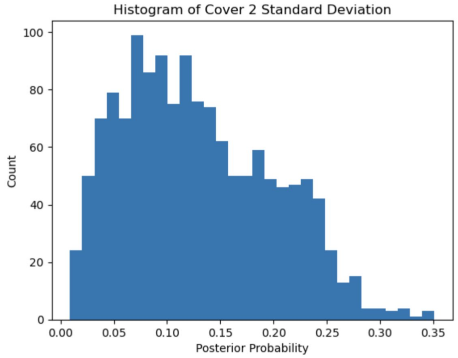
Cover 3





Cover 2



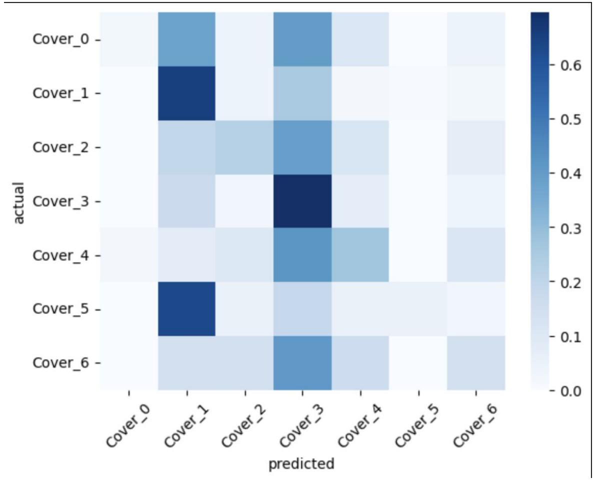


Conclusion

- Certain coverages are more difficult to predict with pre-snap positioning
  - Cover 1 and Cover 3 by far most commonly predicted coverages
- This is intentional - NFL teams attempt to disguise coverages to confuse offense
- Incorporating post-snap player movement would likely improve model
  - Needs to achieve significant improvement in accuracy to replace manual film study

Average Predicted Coverage Probability by Actual Coverage

Actual Coverage



Predicted Coverage

Perfectly performing model would show identity matrix