BAYESIAN CLASSIFICATION FOR NFL DEFENSIVE COVERAGES

Matt Manner, Connor Nickol, Josh Gen

Problem Description:

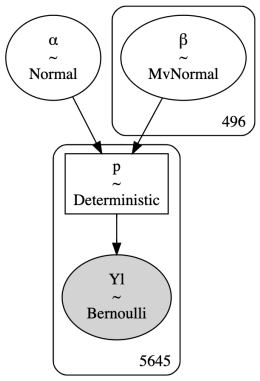
NFL football teams spend enormous resources on studying opposing teams' tendencies and play calls. Given the thin margin for error in the game, the ability for an offense to quickly and accurately predict the defense's coverage can greatly affect outcomes with multi-billion dollar ramifications. Our goal for this project is to use multiple one vs. rest Bayesian logistic regressions to determine the most likely defensive coverage, based on the locations of all 11 defensive players pre-snap, with appropriate uncertainty measures for predictions. Our datasets, which include information such as player location tracking data, play result, game situation, etc, were provided to us by the NFL as part of their "Big Data Bowl" competition. Our model has two potential use cases. The first is aiding NFL quarterbacks in recognizing defensive coverages pre-snap. Quarterbacks can use our model to determine the effect of defensive player positioning pre-snap with the actual coverage run. The second, and more practical use case, is saving thousands of labor hours that team staff currently spend watching film and categorizing coverages. This model seeks to automate this process.

Probability Model:

For our problem, we utilized logistic regression. Each of our regression equations uses a sigmoid transformation of a linear combination of the data and predictor parameters to calculate p_i, the probability of the response variable being 1 for the i^{th} observation.

$$\mu = \beta_0 + \sum_j \beta_j X_j$$

$$p = \frac{1}{1 + e^{\mu}}$$

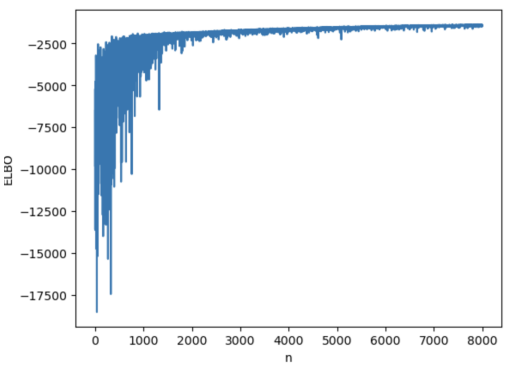


Approach:

There are multiple viable Bayesian methods for a classification problem of this nature, including Naive Bayes and Linear/Quadratic Discriminant Analysis. We chose Bayesian logistic regression because we wanted to study the uncertainty of our predictions. With a large number of predictor variables, logistic regression provides the best option for exploring credible intervals on our outcome variable probability. Because we had seven different coverage options, we made seven different one vs. rest logistic regression models with each coverage as the outcome variable. We then took the highest predicted probability for each play and considered that coverage the “predicted coverage”.

To transform a player’s location (originally represented by an x and a y coordinate), we divided the defensive playing area into 496 tessellating boxes. The box in which the player started the play was coded as his location for the play. Then, we converted the boxes into dummy variables, meaning each of the 496 boxes was a variable on each play- the value of the variable corresponded to the number of players located in that box. Because there are only 11 players on defense for a play, the vast majority of box variables were 0 for a given observation. Therefore, the final data set on which we performed our regressions consisted of 496 predictor variables, and a single binary variable corresponding to that regression’s coverage.

We split our data using a train/test split of 80/20 ($N_{train} = 5645, N_{test} = 1412$), then ran variational inference (VI) on our training set to gain posterior distribution estimates for each predictor (priors were Gaussian with $\mu = 0, \sigma = 100$). We chose variational inference due to our large data set, specifically the several hundred predictors. Additionally, because we created seven models, the reduction in computational complexity for VI compared to sampling methods translated to significant time savings.



The ELBO plot on the left shows the convergence between the approximating distribution and the actual posterior over the latent variables. As is visible in the plot, convergence happens around 2000 iterations. This particular plot is for the posterior distributions of the coefficients for the Cover 5 regression.

Results:

To calculate predicted coverage probabilities, we used the parameter posterior distributions acquired from VI to calculate 5000 estimates for each play in the test set. The mean of these 5000 estimates was the predicted probability for a given coverage on a given play, and we also calculated predicted probability standard deviations from this sample as well. Our model was able to predict certain coverages more accurately than others; in particular, the model was relatively successful at predicting “Cover 1” (65.2% accuracy) and “Cover 3” (69.5%), while it correctly predicted Cover 5 only 5.3% of the time. This may be due to Cover 5’s relative rarity in the NFL, and the similar pre-snap alignment to other coverages like Cover 2 and Cover 4. The model’s accuracy in predicting Cover 1 and Cover 3 may be partially explained by their ubiquity in NFL defensive play calling - those two coverages alone accounted for 58.1% of the observations.

The confusion matrix in Appendix C reveals which actual coverages were predicted incorrectly and the predicted coverage in these cases. For example, Cover 1 and Cover 3 have similar pre-snap defensive alignment; both have one defender far from the line of scrimmage. This is referred to as a single high safety. In fact, the only major difference between these coverages is the motion of the defense *after the snap*, something that our model does not incorporate. This explains the model’s frequent confusion between Covers 1 and 3 in the confusion matrix. When the actual coverage was Cover 1, the model incorrectly predicted Cover 3 25% of the time. The reverse happened at a high rate as well (17% of actual Cover 3 plays were incorrectly classified as Cover 1).

The histograms in Appendix A provide insight into the uncertainty of our model. The probability column of histograms reveal the distributions of predicted mean probabilities for each coverage. Cover 0 and Cover 5 seem to have the most significant right skew, indicating that the model predicts these coverages less than all other coverages. These are by far the least common coverages in the data, with both occurring less than 5% of the time in the dataset. It is interesting to note that both of these coverages have low standard deviation, likely a result of the infrequency of these coverages. In other words, the model consistently predicts Cover 0 and Cover 5 with such low probability that there is little variance in the prediction probabilities.

The most powerful results are gleaned from a comparative analysis of the posterior distributions of the more common coverages: Cover 1, 2, 3 and 4. Cover 3 has the most

different mean probability histogram, with a much more even spread compared to the wide right skews of the others. This indicates that the model often predicts higher probabilities of Cover 3 than the other coverages. Cover 3 is distinct from other options with regards to the players' pre-snap alignment, and it is the most common coverage in the data. This is likely underlying the model's higher distribution for Cover 3 mean probability compared to other distributions. The standard deviation histograms also help interpret the uncertainty behind each coverage. Cover 1 and Cover 3 seem to have higher distributions of standard deviation (left skewed), while Cover 2 and 4 have lower distributions (right skewed). This indicates that the overall confidence of Cover 2 and 4 is slightly higher than Cover 1 and 3, even though these coverages are predicted less often.

Conclusion:

We met our objective of providing a baseline predictor for NFL defensive coverages. Overall, our model performed to an accuracy of 47.0%. Expected accuracy from random guessing would be 14.3% (1/7), while guessing Cover 3, the most common coverage, every play, would yield an accuracy of 33.2% based on the actual frequency for each coverage. Therefore, our model provides some real predictive value, but is unable to correctly predict over 50% of play coverages.

However, given the constraints we imposed on our model (namely, only incorporating pre-snap defensive player locations, rather than player movement throughout the entire play), we would not expect high accuracy- football defenses attempt to disguise their coverage pre-snap to confuse the offense and quarterback. This somewhat arbitrary limitation prevents our model from being accurate enough to replace the many hours that coaches and staff members currently spend watching and categorizing film. A useful extension of our model would be to determine which location boxes (predictor variables) are most determinative of different coverages. Overall, while our model's accuracy was below 50%, we see clear steps to improve the predictive performance.

References

"Fast and easy gridding of point data with geopandas," *james-brennan.github.io*, Mar. 16, 2020. https://james-brennan.github.io/posts/fast_gridding_geopandas/

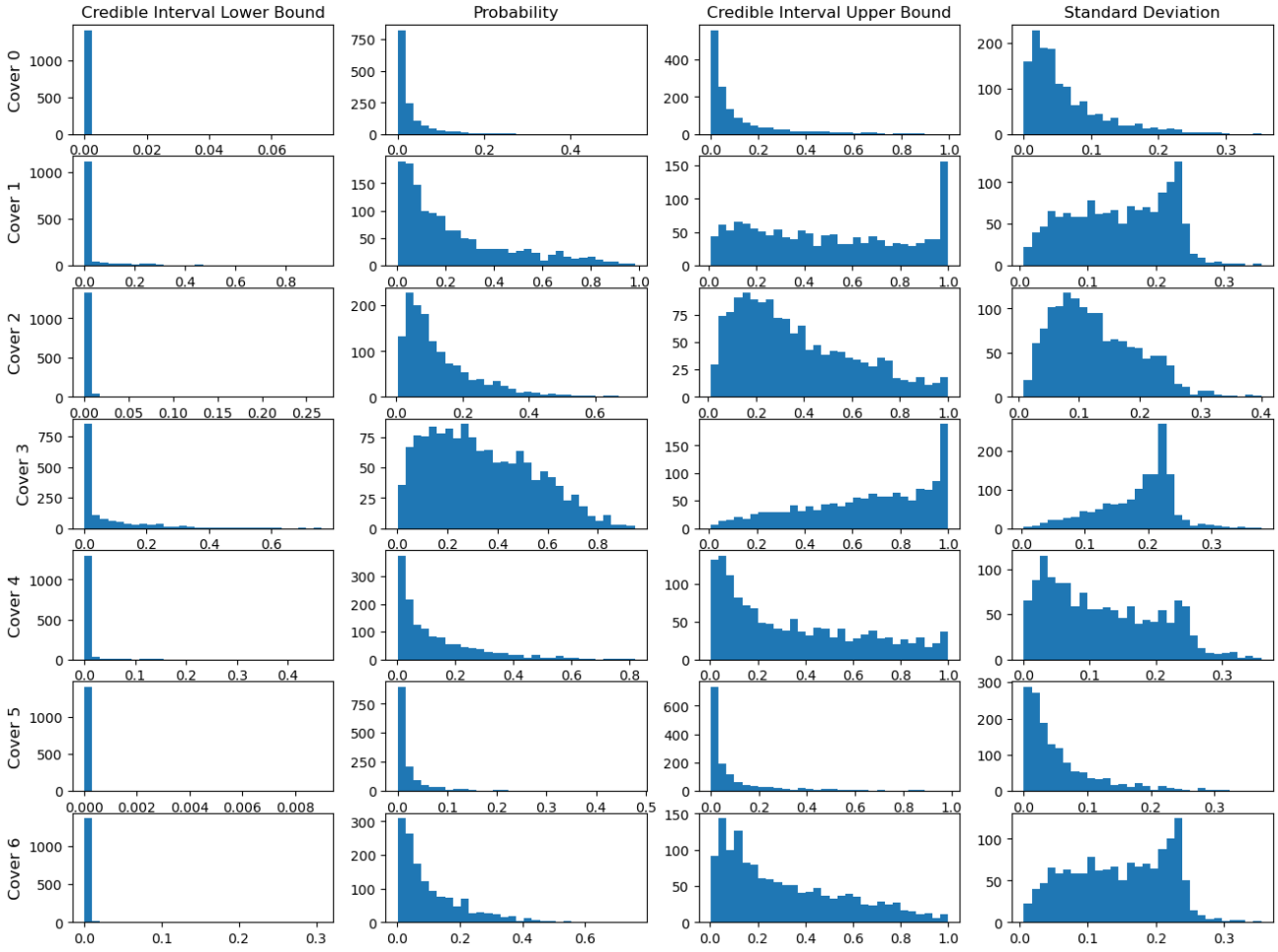
"Scikit-Learn Example in PyMC: Gaussian Process Classifier - Dr. Juan Camilo Orduz," *juanitorduz.github.io*. https://juanitorduz.github.io/sklearn_pymc_classifier/ (accessed Dec. 11, 2022).

Data from: https://www.kaggle.com/competitions/nfl-big-data-bowl-2023

Appendices

Appendix A

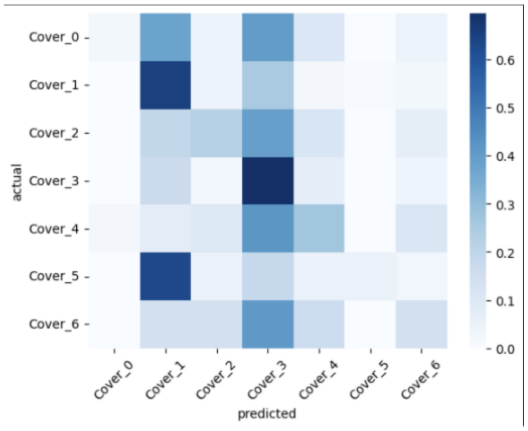
Histograms of Coverage Posteriors



Appendix B

Average Predicted Coverage Probability by Actual Coverage

Actual Coverage



Predicted Coverage

Perfectly performing model would show identity matrix

Appendix C

predicted actual	Cover_0	Cover_1	Cover_2	Cover_3	Cover_4	Cover_5	Cover_6
Cover_0	1	17	2	18	5	0	2
Cover_1	0	219	14	86	7	2	8
Cover_2	0	40	47	82	25	0	15
Cover_3	0	80	14	326	32	0	17
Cover_4	4	14	19	79	51	0	21
Cover_5	0	24	2	7	2	2	1
Cover_6	0	18	18	52	21	0	18