Predicting American Football Games

Alex Crease, Matt Wismer

All code for this post can be found on Github: https://github.com/MattWis/PoissonFootball

Question

We set out to determine how to predict American football games using Bayesian methods. We chose this problem both because of our enjoyment of American football, and the interesting modeling decisions that we would get to make. Football has a lot of complexities; there are 5 different methods of scoring, and strategy determines how the much actual time the 60 game minutes take.

The First Model

To begin, we started by ignoring the uncommon ways to score. We focused on 7 point touchdowns and 3 point field goals. We assumed that every touchdown would come with a one point extra point, and that safeties never happen. We also assumed that scoring happens in a Poisson manner, so that it is equally likely to score at any point in time. We did not incorporate overtime modeling. We found the probabilities of the first team winning, the second team winning, and the game going into OT.

We started out our model by thinking about both types of scoring in football as separate Poisson processes. Prior data was initially taken from net average touchdowns and field goals scored per game. By obtaining distributions of both touchdowns per game and field goals per game based on this data, weighting each distribution and combining them, we obtained a distribution of possible scores in a football match.

In order to improve our model and predict matches, we scraped *covers.com* for data on scores of previous games. We developed specific distributions per team and used those distributions to predict a team's performance during a game as the game progressed. We were then able to compare the probability mass functions and determine the probability that one team would win over another.

To test our model, we decided to predict the October 12th, 2014 game between Matt's hometown Eagles, and Alex's hometown Giants. We felt that focusing our specific efforts towards one game, while keeping our code broad enough to include more games, would be a good way to not get lost in the sea of games that occur on a given weekend in the regular season. This way, we would spend more time working on the model, and less time checking scores.

Predictions (Spoiler Alert: The Eagles Won)

Our first distribution comes from a prior of the overall averages of touchdowns and field goals across the NFL, which is not team specific. We took as our input data the average touchdowns and field goals so far in the 2014 season of each team in the NFL. The results of this are summarized in Figure 1.

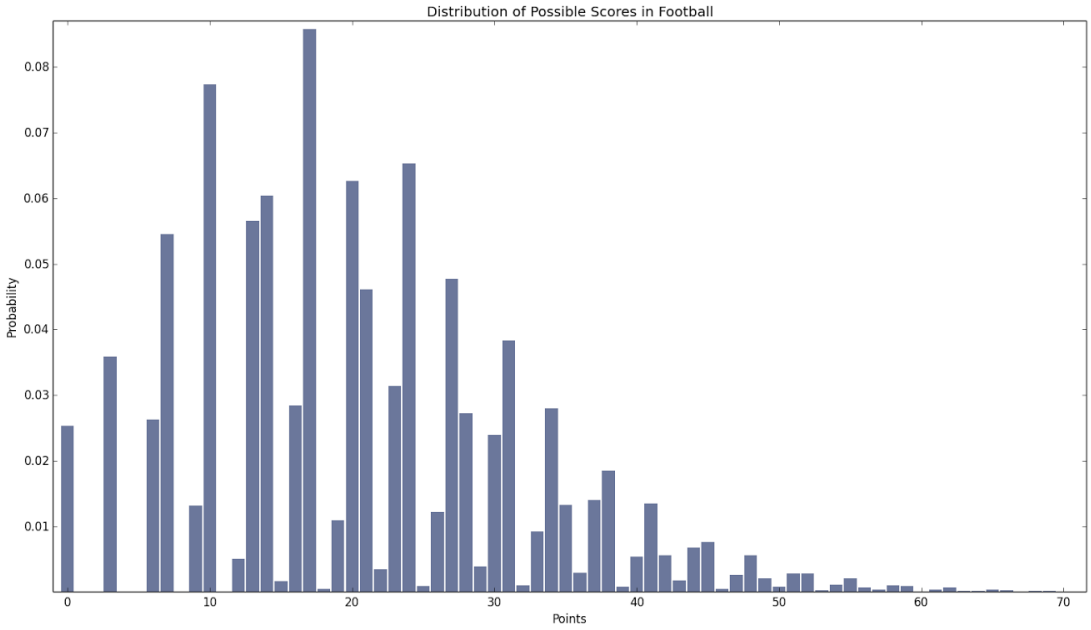

Figure 1: The NFL average distribution of points scored.

Combining the distributions of touchdowns and field goals show that some points are more likely to be scored than others because of the sum of the individual Pmfs. For example, it is impossible to score 1, 2, 4, 5 or 8 points because you cannot create those numbers from multiples of 3 and 7, but 17 is the most probable, with two touchdowns and a field goal, because scoring that amount of each type of goal is fairly likely. However, the data does have the overall trend of a Poisson process, increasing to the maximum and falling off approximately exponentially.

Data from each of the teams' records were used to generate team specific Pmfs, and those were used to generate a prediction on games between two teams. The probability that the Eagles would win was 49.4%, and that the Giants won was 48.2%, with a tie probability of 2.4%, so they are pretty evenly matched, as shown by the cumilative distribution functions in Figure 2.

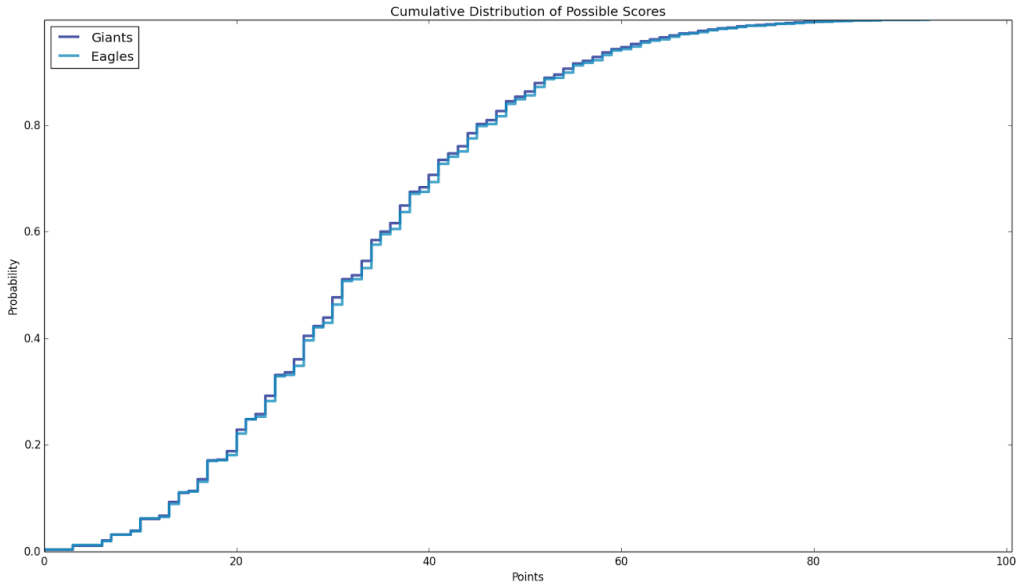

Figure 2: Cdfs of points scored for the Giants and the Eagles.

We updated the probabilities as the game progressed. At half-time, the score was Eagles: 20, Giants: 0. We updated our priors, and the updated Cdfs are shown in Figure 3. The probability that the Eagles would win increased to 90.5%, with the probability of the Giants winning at 8.2%, and a 1.3% chance of a tie. This matches our intuition about the likelihood of a comeback when losing 20-0.

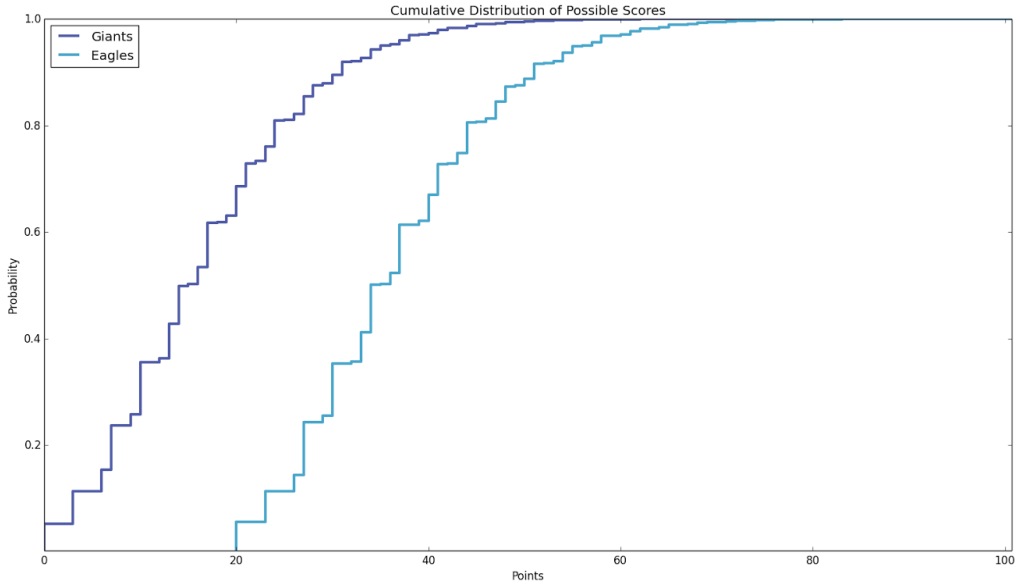

Figure 3: Updated Cdfs of points scored at halftime.

Interpretation and Improvements

We then rethought our model. We were a little worried about the broadness of our original model. The 90% credible interval for the Giants was between 10 and 61. 61 points in a football game is pretty unlikely, and occurs much less than 5% of games. Only 20 regular season games since the 1920s have had more than 60 points scored. (reference) Given that there are 256 regular season games per year, we would expect the probability of scoring 61 points to be closer to 20 / (90 * 256) = 0.08%.

Alternate Model

In light of this disparity, we decided to try modeling the problem a different way. We considered scoring to be a single Poisson process, and have a separate distribution that represents our belief about the probability of a team scoring a touchdown vs a field goal. We could then calculate the probability of scoring a certain number of touchdowns and field goals based on the binomial distribution. We think that this would be a better model because types of scoring are related to each other, in that there will never be a touchdown and a field goal at the same time. So then the single Poisson process is related to the ability of the offense to get into field goal range, and the percentage of touchdowns is related to the team's red zone efficiency.

Predictions from the Alternate Model

This change to the model resulted in the Giants' credible interval moving to between 7 and 55. We were happy about the general tightening, but a little worried that it moved down as a unit. Matt, overall, was pretty happy that the Eagle's credible interval only moved from (10, 62) to (10, 60). This also corresponded to the Eagles moving to a 55% probability of winning, and the Giants down to a 43% probability of winning. The Cdf of our alternate model is shown in Figure 4. The Cdfs have moved apart, which represents that the information known so far about the season is being used more effectively.

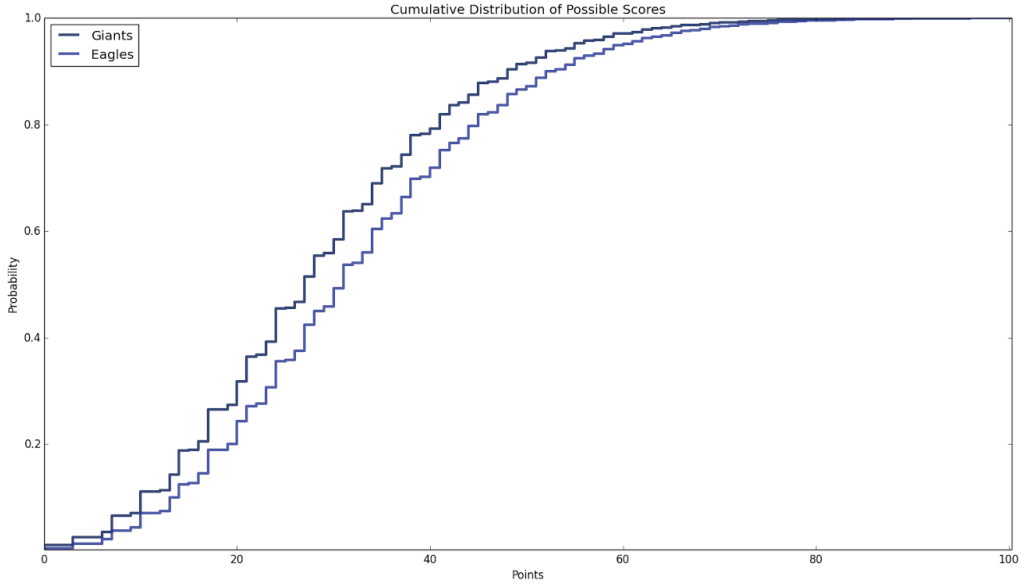

Figure 4: Cdfs of points scored in the alternate model.

We also updated the alternate model with the halftime scores, to see how much it agreed with the first model, in the face of pretty convincing data. The Eagles' probability of victory was updated to 93%, and the Giants' probability was lowered to 6%, with a 1% chance of a tie. While the exact numbers do differ from the first model, the fact that both models moved up to the 90% range when given the same data helps to show that they are both realistic.

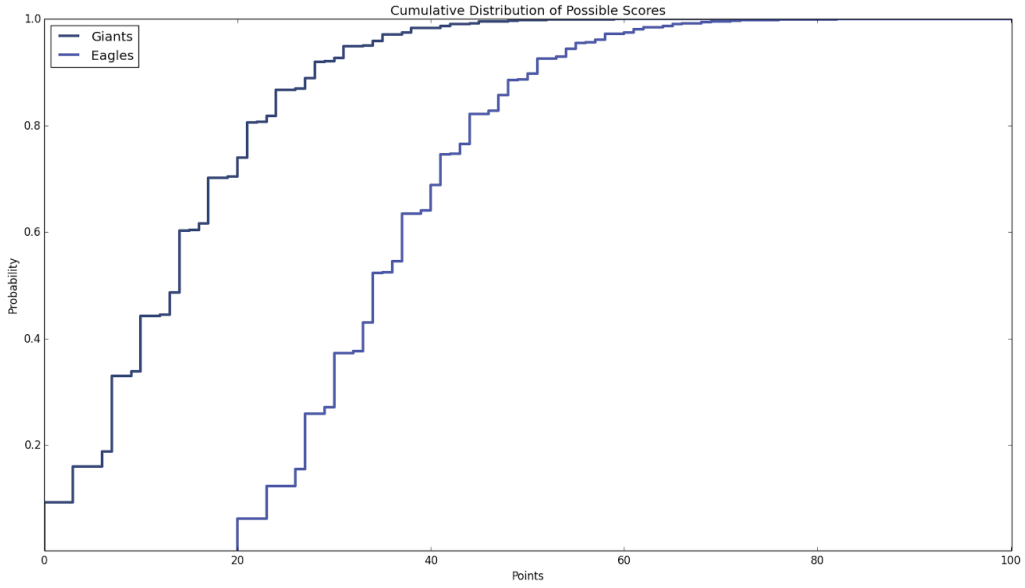

Figure 5: Updated Cdfs with halftime results.

Interpretation

We estimated the probability of a single football game through the course of this project using two separate models. Both models were similar, and it seems that the spread involved was caused by the limited amount of data that was used to construct the distributions. It will be interesting to see if the model gets more accurate as more games are played in the season. Finding a good way to add more data to the model would be a worthwhile extension.

A main limitation of both of these models is the lack of a defense. As Ray Lewis said in the opening of Madden 2005, "Defense wins championships." By only focusing on the points scored by each team, we ignore information about approximately half the players on the team. Incorporating a defense with each team will probably be the next step with this model.