STAT 4800 Final Project

Iain Muir (iam9ez), Hriday Singh (hns4dc), Connor Smith (cms6xs)

Question of Interest

Is there an optimal split between running and pass plays in college football to maximize expected points?

Data Cleaning Process

To prepare the data for our final model, we implemented a variety of data preprocessing steps. First, we removed any plays that took place in "garbage time"—periods of the game in which Pro Football Focus deemed that the outcome of a game had effectively already been decided. Additionally, we created a custom field position variable which we believe intuitively makes more sense that the negative to positive field position variable that already exists; our variable ranges from 1 to 100, with 1 being one yard from your own endzone and 99 being one yard from your opponent's end zone (any positive yards gained increased your position).

Specifically for the field goal Logistic Regression, we began by subsetting the data set to only include instances of field goal attempts. Next, we selected the only two variables we thought were relevant to the kick result, hash and distance. Considering machine learning works more effectively with numeric inputs, we used a One-Hot encoder to create a sparse matrix for the hash variable.

Expected Points Added Model

We created an expected points added function that takes in the inputs of down, distance, and field position and returns a number that indicates the average amount of points a team can be expected to score based on that state. Additionally, we included two extra parameters, run percentage and aggression, to experiment with the impact of play call selection and a coach's risk tolerance.

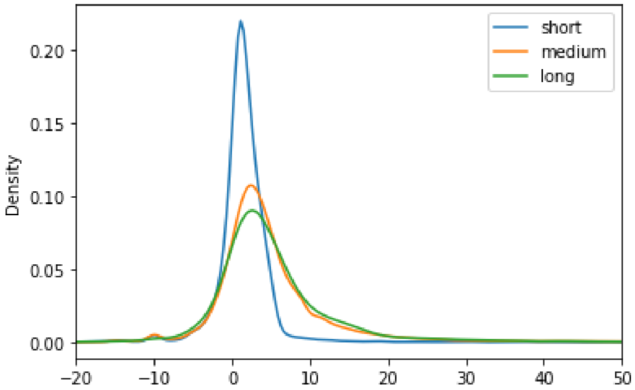
The function effectively simulated a fixed number of scoring periods given an initial state. Within each simulation, a run play function was recursively called until a scoring play was reached. The core of this run play function is an extensive conditional structure that controls which mixture models to sample. Specifically, given the choice of run or pass, the down, and yards to go, a mixture model was indexed from a dictionary of fitted models. The percentage run parameter determines the probability of the coach selecting a run versus pass play, while the aggression parameter determines the coach's decision of 4th down; the higher the aggression, the more likely the coach will go for a 4th down conversion or kick a longer field goal. The following output is an example of three simulations (1st and 10 on your opponent's 20 yard line):

Simulation 1 — 3 points
0 YD PASS, -4 YD RUN, -5 YD RUN, Made 46 YD FG
Simulation 2 — 3 points
0 YD PASS, 0 YD PASS, 0 YD PASS, Made 37 YD FG
Simulation 3 — 3 points
5 YD RUN, 3 YD RUN, 1st Down: 10 YD RUN, -1 YD PASS, -4 YD RUN, 0 YD PASS, Made 24 YD FG
Simulation 4 — 3 points
0 YD RUN, 1 YD PASS, 0 YD PASS, Made 36 YD FG
Simulation 5 — 7 points
0 YD PASS, 1st Down: 16 YD RUN, TOUCHDOWN – 5 YD PASS

In this example, our model determined that the expected points added for this state would be 3.80. Intuitively this makes sense considering the offensive team will almost certainly score at least a field goal.

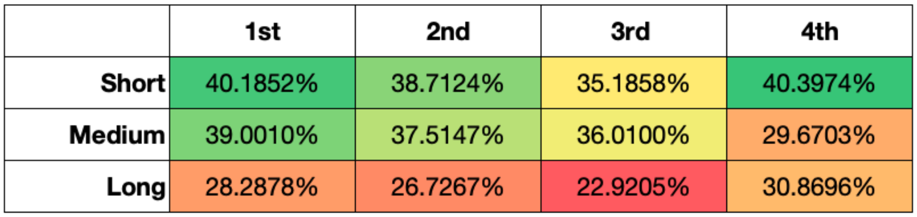
Methodology and Decisions

We decided to use a Bayesian model with the following splits (from top to bottom): a) Whether the play was a run or a pass, b) the down on which the play occurred, and c) the yards to go until first down. From here, we fit Gaussian mixture models to "gain loss net" to obtain the "type" of play (short, medium, or long) distributions for yards gained on each play. We chose the Gaussian mixture model because it was clear that sub groups existed within the overall population and when doing some exploratory data analysis on the yards gained for each play, each sub group appeared to have a distinct unimodal trend although some of the tails were slightly large). An example can be seen here:



Each Gaussian mixture model appeared to have two distinct inflection points which helped us indicate which portion of the distribution each type of play should be sampled from. This fitting occurred for each sub group with different distributions being created and sampled from for each group. The output of this process was a distribution for yards gained in a state defined by [run/pass, down, yards to go, type of play]. This all culminates into a single play.

When each play is run, we also were interested in the probability of success for a certain strategy, particularly for passing (e.g.,a short pass on 1st down with 10 yards to go should be more successful than a long pass on 4th down with 30 yards to go). We implemented this by empirically finding the success rates of each short, medium, and long play from the dataset split by downs. In this case, we used pass yard depth as a proxy for the intended play type (e.g., a play that travels 2 yards in the air but gains 20 yards should be classified as an intended short play rather than a long play). We defined a successful play as one that achieved the median or more net yards gained/lost of the given state and was completed. These successful pass rate values can be seen here:



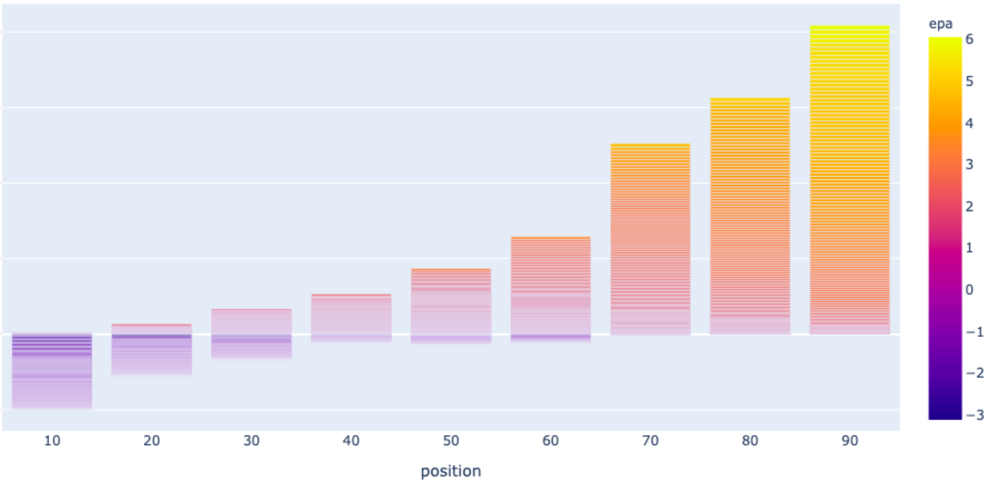
	1st	2nd	3rd	4th
Short	40.1852%	38.7124%	35.1858%	40.3974%
Medium	39.0010%	37.5147%	36.0100%	29.6703%
Long	28.2878%	26.7267%	22.9205%	30.8696%

When implementing this process, a random number generator with the state dependent success rate a parameter is run; if the play results in a success, the sampled yards gained is returned and if the play results in a failure (i.e. incompletion or did not achieve the median value), the play results in 0 yards achieved and the down being incremented.

Field goals were modeled using a logistic regression model with two parameters: distance and hash. We chose this model as we were interested in modeling the success of the field goal with a binary output. Hash did not appear to be an extremely important predictor, but we chose to include it as it did provide some variability and a slightly better model fit.

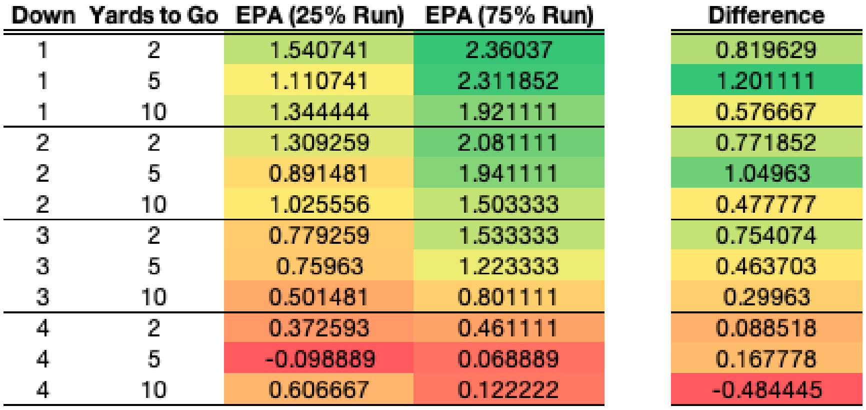
Analysis

Our Expected Points Added model appears to be performing quite well. For evidently advantageous situations, the model outputs a number close to 7, while the model outputs close to -3 for disadvantageous situations. For example, 1st and 10 on your opponents 10 should yield around 6 expected points, while 4 and 10 on your own 10 should likely yield negative; this is because the offensive team will punt and give the opponent good field position. This spectrum of expected points added is visualized below:



As expected, the expected points added varies dramatically depending on field position, and increases as you approach the opposing end zone. Although not clearly shown in the graphic, variation within a bar primarily depicts the effect of down and yards to go; 1st and 2 on the opponents 10 should noticeably vary from 3 and 10 on the opponents 10.

Specifically with regard to our Question of Interest, it is apparent that leaning more on run plays generally results in greater expected points. As shown in the figure below, the expected points when running the ball 75% of the time was almost exclusively higher than that of running the ball 25% of the time. Especially for short yardage situations, run plays dominate pass plays. Although not depicted, a 50/50 split of run and pass plays falls in between these values as expected.



Down	Yards to Go	EPA (25% Run)	EPA (75% Run)	Difference
1	2	1.540741	2.36037	0.819629
1	5	1.110741	2.311852	1.201111
1	10	1.344444	1.921111	0.576667
2	2	1.309259	2.081111	0.771852
2	5	0.891481	1.941111	1.04963
2	10	1.025556	1.503333	0.477777
3	2	0.779259	1.533333	0.754074
3	5	0.75963	1.223333	0.463703
3	10	0.501481	0.801111	0.29963
4	2	0.372593	0.461111	0.088518
4	5	-0.098889	0.068889	0.167778
4	10	0.606667	0.122222	-0.484445

Other Considerations

There were a few pieces of our model that could have been improved with more time and data. One such problem is the intended play versus the actual play. For example, on a designed passing play, a quarterback may decide to keep the ball and run with it. Our model would consider that a running play, while it was intended to be a passing play. Another consideration our model does not take into account is the actual air yards for a pass. In our model, a screen pass that goes for 50 yards would be counted as a long pass, even though the quarterback only threw the ball a few yards down the field.

Additionally, one pitfall of our field goal Logistic Regression model is that it will predict a make for the vast majority of data points; in other words, the negative class error rate is noticeably higher than the positive class. With the data at our disposal, there likely is not too much we can do given the strong imbalance of makes to misses. That being said, gathering more data, sourcing more relevant features (potentially weather conditions), or incorporating generative learning could all work to make the model more robust.

Our model also does not take into account the strength of a team when running the simulation. The model is based on average play data from the PFF dataset. However, if a very strong team (i.e. Alabama) was playing a weak team, the simulation wouldn’t accurately represent what would happen in that game.

We additionally might consider different types of mixture models for the various situations rather than relying on normality assumptions for our predictions (as college football plays tend to be more volatile, and would have more extreme plays than an NFL system due to the nature of team discrepancies, schemes, etc). We used primarily gaussian mixture models, and while it appears they provided a strong fit, there were a few scenarios where different models could have been more useful (for example, on the prior page 3rd and 4th and long situations appeared to predict EPAs that appeared against conventional wisdom and could be improved with different models).

Finally, with either more time or computing power, conducting simulations on a higher scale and for a much larger set of initial states would likely lead to more refined results. We only observed the difference in expected points between 25% and 75% run plays. Analyzing the fluctuation in expected points with percentage increments of 5% or even 1% would have been much more robust but also too computational expensive.

Conclusions

Our group has concluded that a playbook more heavily skewed to run plays is optimal. We specifically analyzed a 75/25, 50/50, and 25/75 run pass split, and the results clearly pointed to higher expected points for high percentages of run calls; for example, especially for short yardage situations, the expected points between a 75/25 and 25/75 run pass split was as much as 1.20 points. This makes sense, because teams that have run the ball effectively have won more throughout the history of football. Running the ball wears down the opposing defense and slows the game down, allowing more clock to run off. This can be especially effective in college football compared to the NFL, as athletes are smaller and less physical than their professional counterparts, so naturally they will tire more quickly. While passing plays are typically more explosive, they are more risky than running the ball due to the risks of interceptions, sacks, and incompletions. The ideal number is most likely somewhere between 50/50 and 75/25 running plays, but we would need to do more simulations to find the optimal playcalling split.

Overall, our group is happy with the performance and results of our Expected Points Added model, and we believe that it could be very impactful for analyzing the state of a game. Furthermore, as detailed in the section above, we see multiple concrete steps that could be taken to continually improve upon the model and make it more robust.

[See EPA_FinalProject.ipynb for Source Code]