4800 Final Project

Writeup

We began our project with the data cleaning process of the 2019 PFF All Plays data set. We first fixed the field position to have the yardage increase from 0 to 100, in perspective of the offensive team, the team with the ball. A new variable was created with this fixed field position. Observations with garbage time were deleted from the dataset, when teams weren't actively trying to score or the play did not matter for the result of the game. The clock was converted to be numeric. Then, the 4th Quarter and the last 4 minutes of the 2nd Quarter were removed from the dataset, as those drives depended too much on time, which was not a variable we wanted to model on. We finally removed plays with a score differential greater than 28 points (4 touchdowns), as teams may have been less inclined to score with such a large point differential. Through this data cleaning process, it was determined we had all the variables we would need from the full dataset, as well as appropriate data for the simulations we would be running to continue onto the next phase of our project.

The next step was to categorize the amount of yards to go in a down to Short, Medium, Long, 10 yards to go, and Longer than 10 yards categories. When we are running our model, and essentially simulating plays in a sandbox environment, we need to be able to appropriately model that play with decisions and results similar to the situation of the plays. Since the yards to go is a crucial state that determines which part of our model should be called upon in a certain play-scenario, it was important to break this down into enough categories that all had significant amounts of data, but also would accurately represent the states being represented. In a similar fashion, we categorized the field position into the Green Zone, Midfield, and Red Zone. Splitting the field into three groups will allow our model to pull data from each section of the field when we begin the execution process later down the line.

With our categorizations complete, we found the number of observations separated by Down, Field Position Category, Yards to Go category, and Run or Pass play. We separated these groupings by number of observations (greater than 30 and less than or equal to 30), so that our sample size was acceptable when creating a distribution from a certain state. We took a specific state and calculated the Gain-Loss from that state, and then shifted all the data to be allowed to fit a gamma distribution (which doesn't allow values of 0 or less). We fit the data to a gamma distribution with the shifted data, and then recorded the shape, rate, and shift for the specific state. This was repeated for each state with greater than 30 observations. At the completion of this process, for states that were modeled on pass plays, we found the completion percentage and stored them.

To deal with the states with less than 30 observations, we needed to fit them to distributions of similar states that had greater than 30 observations. Our exact adoptions are summarized below:

- First Down and 7 to 9 Yards to go adopted the distribution of First Down and 10 Yards to go
- First Down and 4 to 6 Yards to go in the Green Zone adopted the distribution of First Down and 4 to 6 Yards to go in Midfield
- First Down and 1 to 3 Yards to go adopted the distribution of Second Down and 1 to 3 Yards to go
- Any Fourth Downs with fewer than 30 observations adopted the distributions of their respective states at Third Down

We also added a variable that, for a given state, determined the probability whether the play was a pass. This would help determine later when given a certain state, whether to choose to model from the Run or Pass distribution. For this distribution, even though there are some with small counts, after examination, we determined that they are close enough to what we believe to be the accurate probabilities for those specific circumstances in football (i.e. 4th and long should have a very high probability of pass).

Finally, we were prepared to actually simulate some plays and get results. We created the functions that would actually simulate plays until points were scored by either team. An entry into this set of functions constituted a certain state in a football game, whether our team or the other team had possession, the Field Position, the Yards to Go for a Down, the Down, and the current Score Differential. We categorized the state's numeric values of Field Position and Yards to Go into their aforementioned respective categories, and accessed the respective pass probability given the state. If there was a 4th down state that did not have any prior observations to provide a distribution or pass probability, it adopted its respective 3rd down state's distribution and pass probability. With the given pass probability, we sampled to output whether the decision should be to perform a Pass or Run play. We then stored this decision. Given this state of Down, Field Position, Yards to Go, and Pass or Run play, we then had an available distribution to model on (under the play function).

Nearing the end of our project came the bulk of our work with the set_of_downs function. This is the main function of our EP model, and it is a recursive function that simulates a set of downs and continues until a score is reached, then outputs that score (+ for our team, - for opponents). Within this function, we needed to simulate how fourth down decisions would occur, as the rest of the function was running play until a 4th down came, a score occurred, or the drive continued with a first down (in which case it recursively called set_of_downs with the new inputs). For the fourth down decisions, we investigated how the college games were played at scale, when teams settled for field goals, when teams punted, and when they decided to go-for-it. We then wrote a fourth-down decision function to input into our set_of_down function for 4th downs where if there was less than 2 yards to go, and the offensive team was across midfield, they would go for the conversion. Because our model is expected points and not win-probability, worrying about when it would be beneficial to go for a field goal rather than trying to convert was not factored into our model. The next decision was that if a team was past the 73-yard line, they would elect to kick a field goal. This means if it is more than 4th and two, and the field goal length is less than 50 yards, a team would kick the field goal. We created a logistic regression to estimate the probability of making a field goal given distance of the field goal from the data and sampled on this probability for the specific situation to determine whether the field goal was made or not. Finally, if the fourth down states did not fit into either of these categories, the team would punt.

With a completed set_of_downs function, it was time to investigate our question for the project, grouping into final matrices, and observing the results. We all know how costly penalties are in the game of football, but we specifically wanted to look into the difference in expected points after a holding penalty (loss of 10 yards and redo down) would be from many different spots on the field. This has practical applications for coaches to show their players how much a holding penalty can impact the game, but more importantly for announcers to bring new statistics about the harm of a penalty to the viewers.

As for our deliverable, we wanted to create a visual for the quantitative cost of penalties in certain situations. We decided that the best way to show this was to create three matrices, one from the green zone (starting at the 25 yard line), one from midfield (starting at the 50 yard line), and one from the red zone (starting at the 75 yard line). Each matrix has downs 1-4 as rows and specific yards to go on the columns. The matrices are shown below in the results section of this writeup. Each value in the matrix represents the change in expected points resulting from a holding penalty.

Results

Own 25 Yard Line

		Yards to Go				
		2	5	8	10	15
DOWN	1st	-0.6596	-0.958	-0.8142	-0.7713	-0.6757
	2nd	-0.7322	-0.6746	-0.6341	-0.675	-0.7677
	3rd	-0.7433	-0.5084	-0.4694	-0.7269	-0.4818
	4th	-0.4105	-0.3688	-0.329	-0.3687	-0.3762

50 Yard Line

		Yards to Go				
		2	5	8	10	15
DOWN	1st	-0.8195	-0.9007	-0.8308	-0.7988	-0.6395
	2nd	-0.8313	-0.7325	-0.7267	-0.5896	-0.488
	3rd	-0.7119	-0.4184	-0.4255	-0.2682	-0.1788
	4th	0.2465	0.3293	0.3389	0.3193	0.2289

Opponent’s 25 Yard Line

		Yards to Go				
		2	5	8	10	15
DOWN	1st	-1.2832	-1.4212	-1.5064	-1.6335	-1.5685
	2nd	-2.0513	-1.9814	-1.963	-1.8484	-1.7346
	3rd	-2.2716	-1.8311	-2.091	-1.6689	-1.7296
	4th	-1.9776	-2.0079	-2.1806	-1.9101	-2.0456

There are some interesting interpretations from our matrices above. The first insight that appears from the matrices is that the strongest negative expected points row comes from fourth downs on the opponents 75 yard line. In fourth down situations, this can be the difference between a successful field goal and a missed field goal or the opportunity to go-for-it or kick the field goal. In critical fourth down situations close to the red zones, this area of the matrix demonstrates just how costly a holding penalty can be (almost uniformly -2 expected points).

The most negative cell is 3rd and 2 from the opponent’s 25 yard line. A holding penalty in this situation is extremely costly. First of all, teams are very likely to convert on 3rd and 2 for a first down within the 23 yard line. This is likely to result in at least a field goal. However, a holding penalty would push the team back to 3rd and 12 from the opponent’s 35 yard line. A 3rd and 12 is unlikely to convert for a first down and the team is now essentially out of field goal range as a field goal from the 35 yard line is a 52 yard field goal which is a very long field goal in college football.

Another interesting takeaway is that on your own 25 yard line, a holding penalty on 4th down has much less cost than holding on other downs. This makes sense, as teams are most likely punting from their own 25 regardless of whether it is 4th and 5 or 4th and 15.

The cells that are all green for the 50 yard line fourth downs occur because these are situations where either way the team is punting, so it should not influence the game significantly. In certain situations, we see teams intentionally get delays of game from this territory because of how little the extra penalty yards matter because of the punting. The way we modeled our punting, on any play in this region the team was punting regardless of the penalty, and each time for a touchback.