Uncertainty Quantification in 4th Down Decision Making

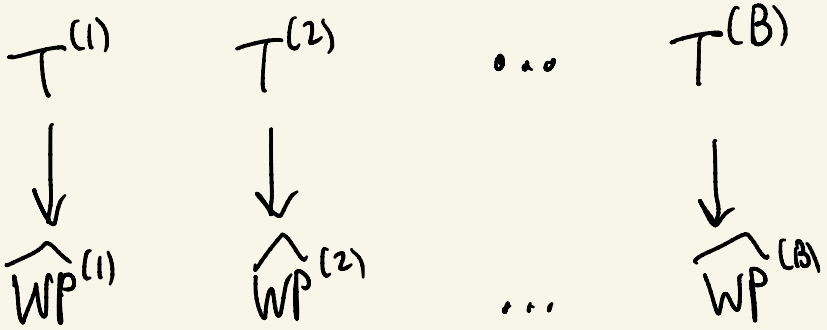
* Use Randomized Cluster Bootstrap to obtain WP CI with adequate coverage.

* Verified the Kosher-ness of this method using a simplified football simulation in which true WP is known.

* Today:

1. Apply RCB to real football data
2. Propagate the uncertainty to the 4th down decision itself
3. Use 4th app with uncertainty quantification

$T = (X, y) \longrightarrow \widehat{WP}$



game-state x. Re-order $1, 2, \ldots, B$ so that

$\widehat{WP}^{(1)}(x) \leq \widehat{WP}^{(2)}(x) \leq \ldots \leq \widehat{WP}^{(B)}(x)$

CI $B = 100$, $\left[\widehat{WP}^{(3)}(x), \widehat{WP}^{97}(x)\right]$

$\widehat{WP}(x)$

We don't really care about CI for $\widehat{WP}$ themselves. We care about how this uncertainty impacts 4th down decision making.

Before $T = (x, y)$

$\downarrow$

$\widehat{WP}_1$

$\downarrow$

$\widehat{WP}_{go}(x)$, $\widehat{WP}_{fg}(x)$, $\widehat{WP}_{punt}(x)$

Say $\widehat{WP}_{go}(x) \geq \widehat{WP}_{fg}(x) \geq \widehat{WP}_{punt}(x)$

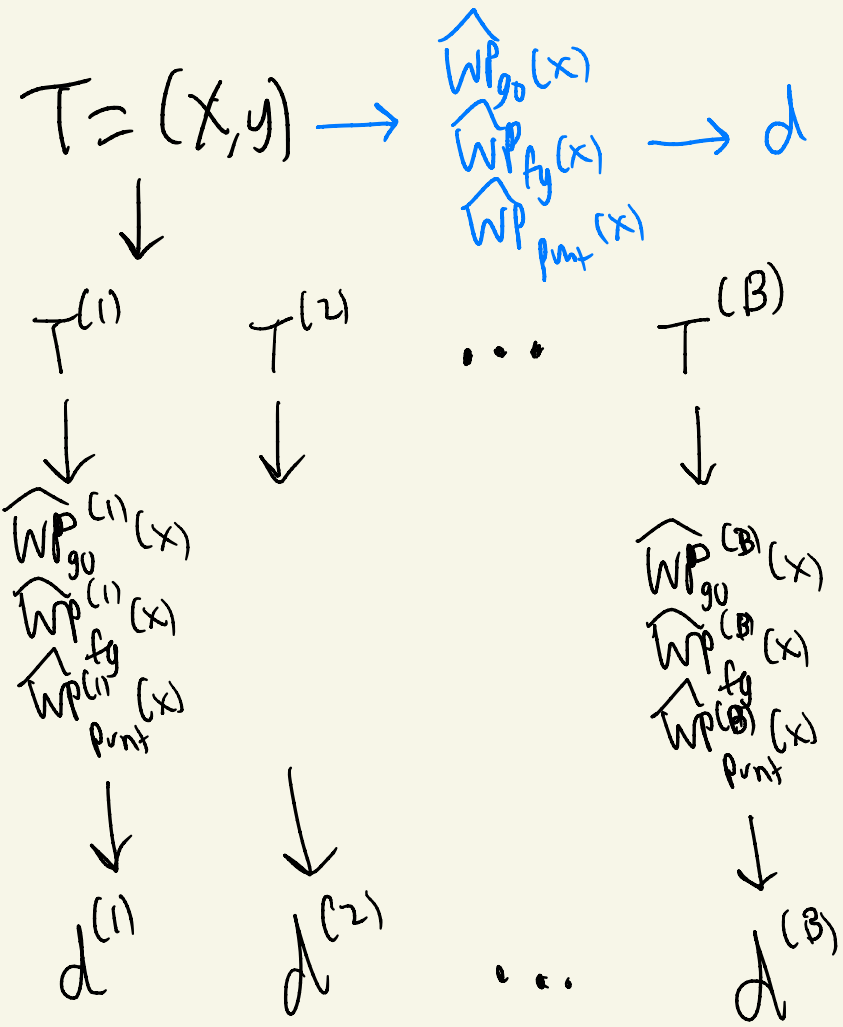
$\longrightarrow$ therefore, go for it

diff $\widehat{WP}_{go}(x) - \widehat{WP}_{fg}(x)$ } by how "confident" they are

Except: what if these estimates suck?

$\rightarrow$ how to incorporate uncertainty quantification into this decision making?

Idea Bootstrap the decision itself rather than the $\widehat{WP}$.



B bootstrapped decisions $\hat{d}^{(1)}, \ldots, \hat{d}^{(B)}$

Original estimated optimal decision $\hat{d}$

decision confidence := bootstrap percentage

$\hat{d} = \text{go}$

our confidence in this estimated optimal decision := proportion of bootstrapped models which yielded estimated optimal decision Go

$\hat{d}^{(1)}, \ldots, \hat{d}^{(B)} \longrightarrow$ go boot. $q_b =$ $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ fg boot. $q_b =$ $\begin{pmatrix} 1/2 \\ 1/2 \\ 0 \end{pmatrix}$ punt boot. $q_b =$

1. estimated optimal decision $\hat{d}$ → effect size $\widehat{WP}_{go}(x) - \widehat{WP}_{fg}(x)$
2. bootstrap percentages $(q_{go}, q_{fg}, q_{punt})$ → decision uncertainty

4.2 Example plays: better fourth down decision making

Our improved fourth down decision making procedure relies on both bootstrap percentage and estimated gain in win probability, which we illustrate using more example plays.

Example play 4. Figure 10 visualizes our decision making for a fourth down play in which the Commanders have the ball against the Colts in Week 8 of 2022. Punt provides a slight edge over Go according to the WP point estimate (+0.004 WP), and 100% of bootstrapped models find that Punt is the best decision. Additionally, our confidence interval of the estimated gain in win

[21] This figure was taken from Burke's Twitter `@bburkeESPN`.

probability by punting is $[0.33\%, 4.64\%]$, which is strictly positive. Thus, we are confident in this edge, even if it is small, and recommend that the Commanders should Punt.

Up 1, 4th & 5, 71 yards from opponent endzone

Qtr 3, 5:53 | Timeouts: Off 3, Def 3 | Point Spread: 3

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	baseline coach %
Punt	0.440	[0.00328, 0.04635]	1				0.934
Go for it	0.436		0	0.426	0.345	0.557	0.066
Field goal	0.345		0	0.000	0.345	0.548	0.000

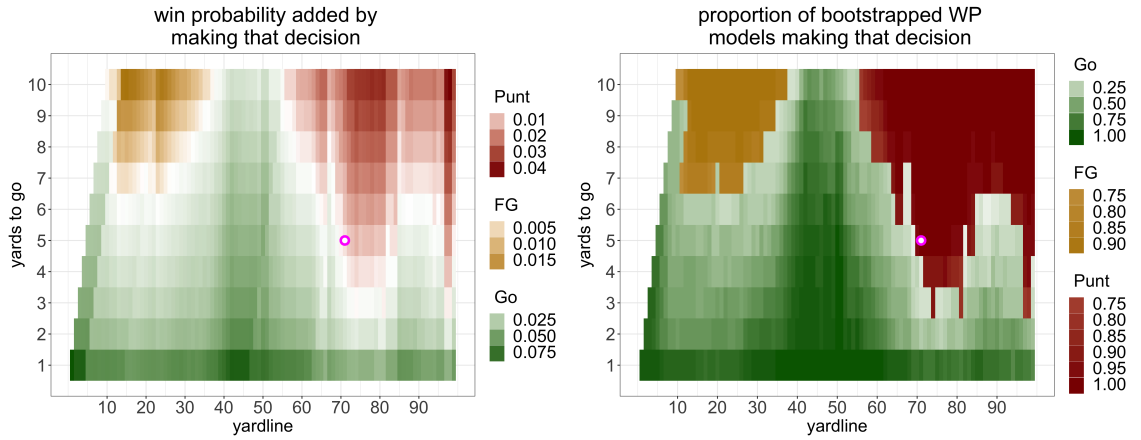


Figure 10: Our decision making for example play 4.

Example play 5. Figure 11 visualizes our decision making for an infamous fourth down play in which the Raiders have the ball against the Rams in Week 14 of 2022. Go provides a strong edge over Punt according to the WP point estimate (4.1% WP), and 96.2% of bootstrapped models find that Go is the best decision. Additionally, our confidence interval of the estimated gain in win probability by going for it is $[0.30\%, 5.23\%]$, which is strictly positive. Thus, we are confident in this edge, and we recommend that the Raiders should Go.[22] We recommend this decision much more strongly than we recommend the previous play's decision, even though they have similar bootstrap percentage, because the WP point estimate is so much larger.

Example play 6. Figure 12 visualizes our decision making for a fourth down play in which the Bears have the ball against the Jets in Week 12 of 2022. FG provides a solid edge over Go according to the WP point estimate (1.8% WP), but 38.5% of bootstrapped models find that FG is the best decision.[23] In other words, we don't have enough data to believe it is the best decision; the data is not confident in its own point estimate. Moreover, our confidence interval of the estimated gain in

[22] In real life, the Raiders punted.

[23] Also, note that the (yardline, yards to go) point is far from the decision boundary, but that doesn't imply anything about the decision uncertainty.

Up 6, 4th & 1, 66 yards from opponent endzone

Qtr 4, 2:00 | Timeouts: Off 2, Def 0 | Point Spread: -6.5

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	baseline coach %
Go for it	0.927	[0.00295, 0.05236]	0.962	0.691	0.783	0.991	0.229
Punt	0.886		0.038				0.771
Field goal	0.783		0.000	0.000	0.783	0.980	0.000

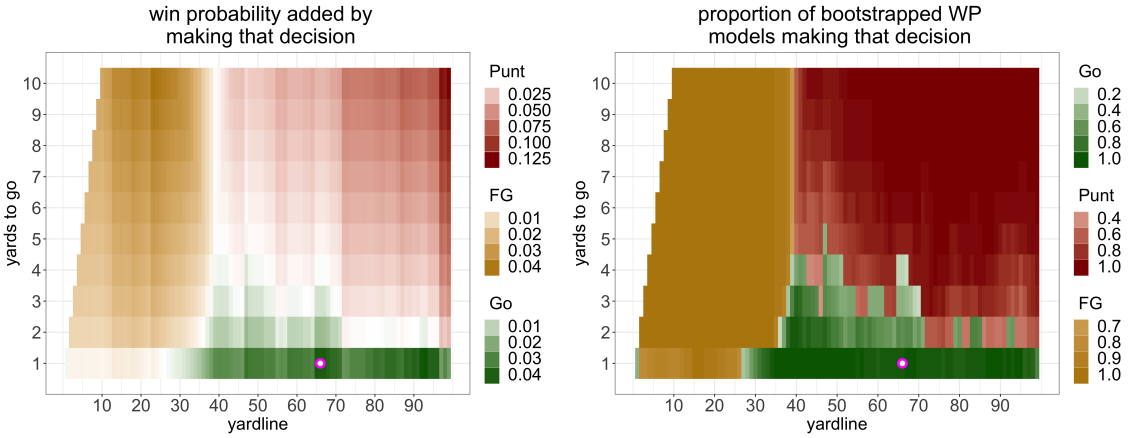


Figure 11: Our decision making for example play 5.

win probability by going for it is $[-3.99\%, 4.26\%]$. This reflects that FG could either be a great or a terrible decision. Therefore, we suggest that a football team should use some other method to pick between kicking a field goal and going for it. For example, in such a situation of high uncertainty, a coach's gut (or internal model) may be better than the edge implied by WP estimates. The coach spends a significant amount of time with his players, and he may notice information which doesn't show up in the data. For instance, if Bears coach Matt Eberflus notices that kicker Cairo Santos is particularly hot today and quarterback Justin Fields appears a bit lethargic today, perhaps Eberflus should be able to choose to kick a long field goal without being ridiculed. On this view, we should evaluate a coach's fourth down decision making on plays where the bootstrap percentage according to our model is high (say, a bootstrap percentage above 85%).

Example play 7. Figure 13 visualizes our decision making for a fourth down play in which the Eagles have the ball against the Chiefs in the the 2023 Super Bowl. Go provides a solid edge over Punt according to the WP point estimate (2.7% WP). But, 76.9% of bootstrapped models find that Go is the best decision, and our confidence interval of the estimated gain in win probability by going for it is $[-2.9\%, 4.83\%]$. This bootstrap percentage is high enough where we lean towards Go as the better decision, but the confidence interval suggests that it is still possible that Go is a terrible decision; we don't have enough data to know. On this view, even if we lean towards Go, the data isn't speaking strongly enough to overrule a coach who may notice subtleties such as

Down 7, 4th & 4, 4 yards from opponent endzone

Qtr 1, 6:02 | Timeouts: Off 3, Def 3 | Point Spread: 8.5

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	baseline coach %
Field goal	0.183	[-0.03993, 0.04263]	0.385	0.987	0.11	0.184	0.849
Go for it	0.165		0.615	0.467	0.11	0.228	0.151
Punt	0.099		0.000				0.000

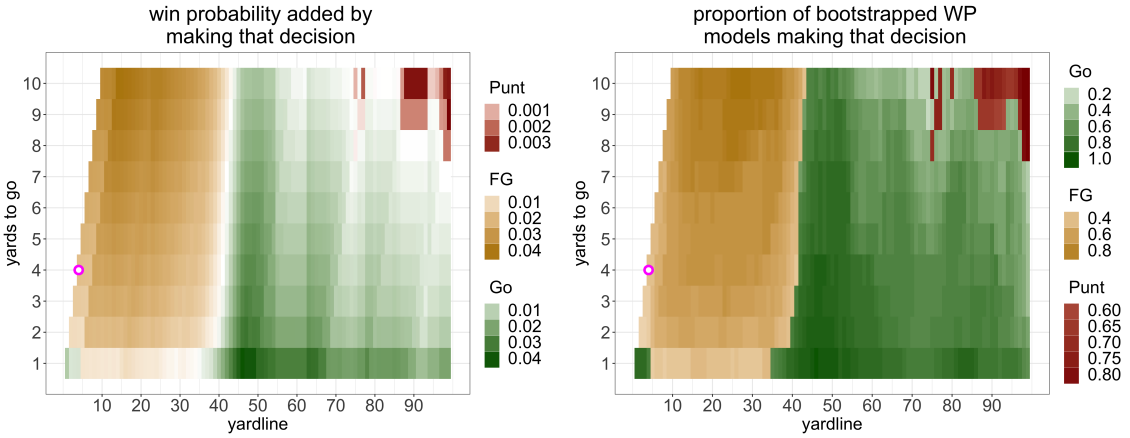


Figure 12: Our decision making for example play 6.

momentum or hotness in his players on a given day.

4.3 Analytics, have some humility

Before, fourth down decision making used estimated win probability gain as the basis of decision making. We extend this decision making procedure to include uncertainty quantification because win probability estimates come from a statistical model fit from observational data. In particular, we quantify decision uncertainty by bootstrapping the decision itself. We find that that far fewer fourth down decision are as obvious as analysts claim. Models could be biased, wrong, missing covariates, and overfit; and even if the model is right, for a huge proportion of game-states there is not enough data to be confident in win probability point estimates. After all, there have only been about four thousand games in the last fifteen years. Therefore, we suggest that football analysts have more humility and accept the limitations which result from having limited data.

Down 1, 4th & 3, 68 yards from opponent endzone

Qtr 4, 10:00 | Timeouts: Off 2, Def 2 | Point Spread: -1.5

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed	baseline coach %
Go for it	0.420	[-0.02904, 0.04827]	0.769	0.498	0.301	0.539	0.12
Punt	0.393		0.231				0.88
Field goal	0.301		0.000	0.000	0.301	0.618	0.00

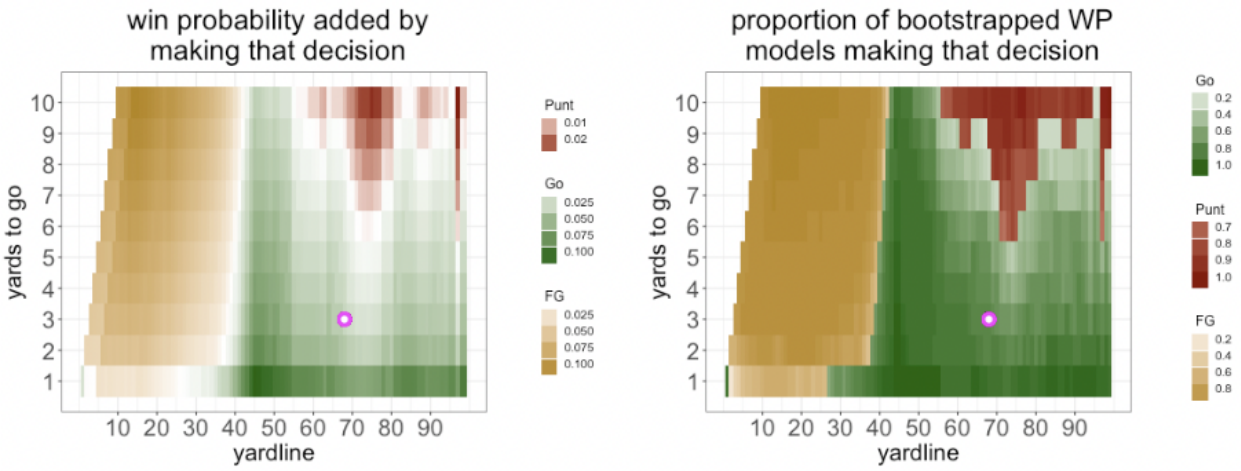


Figure 13: Our decision making for example play 7.