Nonparametric Bootstrap Uncertainty Quantification

3.4 Applying our win probability methods to real football data

We fit a win probability model fit from historical American football data using [redacted] XGBoost [redacted] Additionally, as justified by our simulation study, we use a randomized cluster bootstrap to obtain approximate WP confidence intervals. We resample $G/2$ games in each of $B = 26$ bootstrapped datasets, where G is the number of games in our dataset. As before, we widen the confidence intervals at the extremes: when $\widehat{\mathsf{WP}} < 0.025$ we make 0 the lower bound and when $\widehat{\mathsf{WP}} > 0.975$ we make 1 the upper bound.

4 Fourth down decision making

Win probability estimates fit from highly autocorrelated historical data, which form the foundation of fourth down decision making, are subject to considerable uncertainty. In this Section, we modify the existing fourth down decision procedure to account for this uncertainty. We find that decision making changes substantially. In particular, we find that that far fewer fourth down decision are as obvious as analysts claim.

Existing fourth down decision making procedure. The existing decision process involves making the decision which maximizes estimated win probability. On this view, existing decision procedures are based solely on *effect size* (e.g., the estimated gain in win probability by making a decision). We discuss how to compute the win probability if a team goes for it, kicks a field goal, and punts in Appendix S3.1.

Our fourth down decision making procedure. The problem with basing a fourth down decision on estimated effect size is that it ignores the uncertainty inherent in estimating win probability from highly autocorrelated observational data. Therefore, we create a decision procedure, justified by the simulation study from Section 3.3, which accounts for uncertainty. In particular, we use a randomized cluster bootstrap to create $B = 26$ bootstrapped datasets. In creating each bootstrapped dataset, we resample with replacement half as many games as in the original observed dataset. Then, within each resampled game, we resample plays with replacement (within each game, we resample the same number of plays as observed in that game). Then, we fit a win probability model using catalytic machine learning to each bootstrapped dataset. Each bootstrapped model produces an optimal decision in $\{\mathsf{Go}, \mathsf{FG}, \mathsf{Punt}\}$, which maximizes the estimated gain in win probability by making that decision. Then, we define the *bootstrap percentage* of a decision $\mathscr{d}$ as the percentage

of bootstrapped models which report $\mathscr{d}$ as the optimal decision. This is the right way to quantify uncertainty in the decision making process because, as decision makers, football analysts should care more about the uncertainty of the *decision itself* than the uncertainty in the win probability estimates of each individual decision[16](Go, FG, or Punt) (Friedman et al., 1999).

Understanding bootstrap percentage. We view the bootstrap percentage as a measure of *data reliability*. Specifically, at each game-state the model produces a point estimate of the fourth down decision; bootstrap percentage tells us how reliable this estimate is, or how much the data trusts its own estimate. To understand, think of the outcome (winning team) of each row (play) in the dataset as a random draw. If some of these draws resulted in different outcomes, our fitted win probability functions would be different. The less data we have access to, the more sensitive models are to the random idiosyncrasies of any particular training dataset. The bootstrap quantifies this sensitivity: given the amount of data we have, it quantifies the spectrum of variability in potential resulting fitted models. To understand, think of our one observed dataset as sampled from some underlying "true" win probability distribution. Then, each bootstrap resampled dataset from this observed sample approximates another sample from the underlying "true" distribution (Efron, 1979). In other words, by creating B resampled datasets, the bootstrap approximates B different "universes" of random outcomes, and it measures the variability of fitted WP functions across these universes. On this view, the bootstrap percentage of a decision $\mathscr{d}$ approximately measures: given the amount of data we have, in what proportion of universes would $\mathscr{d}$ be the optimal decision according to win probability point estimates fit from observed data? For example, suppose 60% of bootstrapped WP models imply Go is optimal and 40% imply FG is optimal. In that case, the decision implied by our estimated WP model fit from observed data is unreliable, for if we re-drew the random outcomes of the winning teams in our dataset, the estimated optimal decision would be differ approximately 40% of the time. In general, the closer the bootstrap percentages of two decisions are, the less reliable the data is in telling us which decision is better.

Although bootstrap percentage and estimated gain in win probability by making a decision are correlated, they are not the same. As we discuss in the next Section, there are myriad examples of plays in which the estimated win probability gain by making a decision is high but the bootstrap percentage is low. For these game-states, we don't have enough data to be confident in the decision suggested by the WP point estimates. Other plays, however, feature a high bootstrap percentage and a high estimated gain in win probability. For these game-states, we do have enough data to be

[16]Uncertainty on this decision itself, which is based on the uncertainty of the win probability *gain* by making that decision, is distinct from uncertainty in the WP estimates of each individual decision because the decision may be correlated across datasets. In other words, even if different random draws of the outcomes of game winners in an observational dataset of football plays yields different estimates of the WP of Go and FG, it may yield similar estimates of the *difference* in WP between Go and FG.

confident in the decision suggested by the WP point estimates.

4.1 Example plays: how fourth down decision making changes

Fourth down decision making changes substantially when we use bootstrap percentage to account for uncertainty in win probability estimates, which we illustrate using example plays.[17]

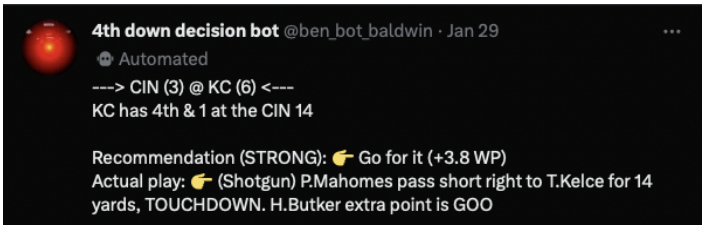
Example play 1. We begin by comparing Baldwin's fourth down decision making procedure to ours. For Baldwin, the strength of a decision is based on the estimated win probability gain by making that decision. For example, Figure 6 shows Baldwin's fourth down decision procedure for a play from the 2023 AFC Championship game in which the Chiefs have the ball against the Bengals.[18] According to Baldwin's model, the Lions have a 72% WP if they go for it and a 68% WP if they punt. Baldwin views the estimated 3.8% WP gain by going for it as a "strong" decision.

Up 3, 4th & 1, 14 yards from opponent end zone
Qtr 2, 03:58 | Timeouts: Off 0, Def 3

			Win % if	
	Win %	Success %[1]	Fail	Succeed
Go for it	**72**	68	64	76
Field goal attempt	**68**	94	62	69

[1] Likelihood of converting on 4th down or of making field goal
Source: @ben_bot_baldwin

(a)



(b)

Figure 6: Baldwin's decision making for example play 1.

But, a decision with a high estimated win probability gain is not necessarily the best decision with certainty. In particular, according to our model, even though going for it produces a positive estimated win probability gain, 53.8% of our bootstrapped models indicate that going for it is better while the other 46.2% indicate that kicking a field goal is better (see the orange column of Figure 7). This reflects considerable uncertainty in the optimal fourth down decision. In other words, we don't have enough data to know with high confidence which decision is better. Moreover, our confidence interval of the estimated gain in win probability by going for it is $[-3.59\%, 4.51\%]$. This reflects that Go could either be a great or a terrible decision.

Example play 2. Next, we compare Burke's fourth down decision making procedure to ours. We begin with Burke's fourth down decision boundary charts. Burke's chart[19] in Figure 8a plots, for various values of yardline and yards to go while holding the other variables constant, the best

[17] To compare our decision making procedure to the decisions that actual football coaches tend to make, we model the probability that a coach chooses a decision in $\{\mathsf{Go}, \mathsf{FG}, \mathsf{Punt}\}$ as a function of game-state. We discuss this *baseline coach model* in detail in Appendix S3.5.

[18] These figure were taken from Baldwin's fourth down Twitter bot `@ben_bot_baldwin`.

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

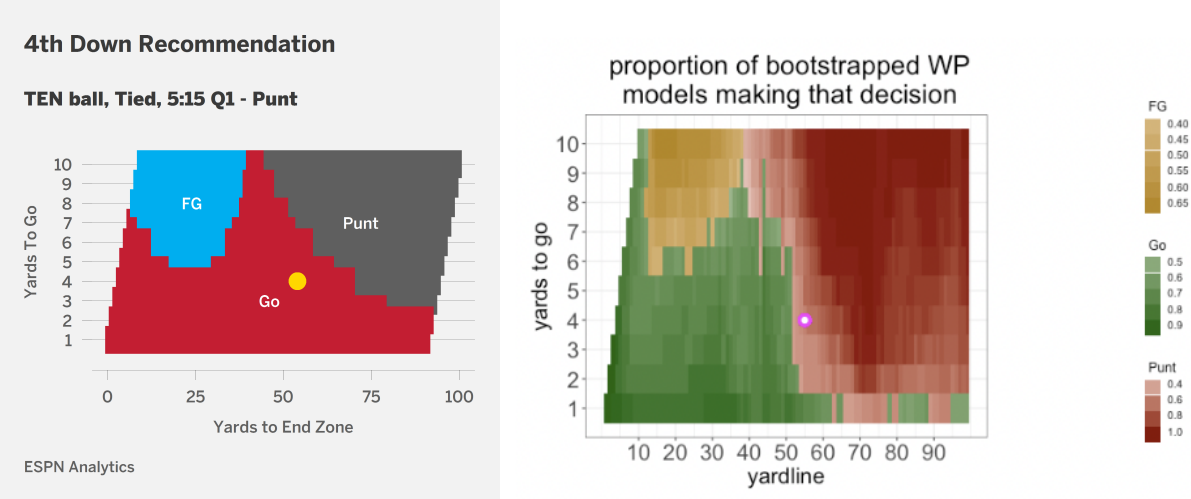
Up 3, 4th & 1, 14 yards from opponent endzone

Qtr 2, 4:00 | Timeouts: Off 0, Def 3 | Point Spread: -2

decision	WP	WP gain CI	boot %	success prob	WP if fail	WP if succeed
Go for it	0.737	[-0.03588, 0.04514]	0.538	0.667	0.667	0.772
Field goal	0.722		0.462	0.923	0.667	0.726

Figure 7: Our decision making for example play 1.

decision as determined by the estimated gain in win probability by making that decision. Burke uses solid colors to indicate which decision to make, and argues that strong decisions are located far from the decision boundary. While it is true that estimated win probability gain is correlated with distance from the decision boundary, a decision that is far from the boundary is not necessarily the best decision with certainty. To illustrate this point, we create Figure 8b which plots bootstrap percentage for various combinations of yardline and yards to go. The color indicates the decision with the highest estimated win probability gain by making that decision.[20] The color intensity indicates the bootstrap percentage (which increases as the color darkens). We see in Figure 8b that there is a substantial zone of decision uncertainty – a massive *S* shape of light color where the bootstrap percentage is below, say, 70% – through the middle of plot. For these game-states, we don't have enough data to know which decision is best.



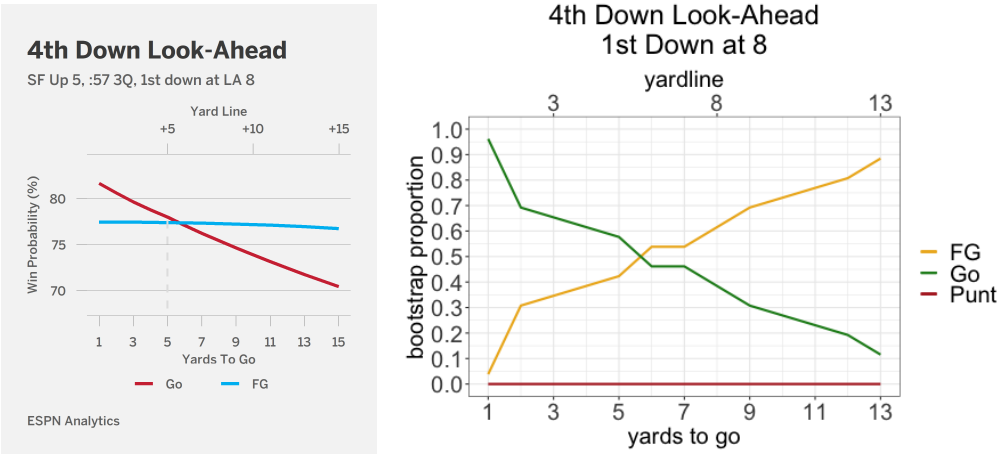
(a) (b)

Figure 8: Left: Burke's decision making for example play 2. Right: ours.

Example play 3. Now, we compare Burke's "fourth down look-ahead" charts to ours, shown in Figure 9. In this play, SF has the ball at the 8 yardline. Burke's chart[21] in Figure 9a looks ahead to the possibility that three plays in the future, SF has a fourth down. If SF ends up with a fourth down

[20] Green is Go, yellow is FG, and red is Punt. These colors come from Burke's thought that fourth down decision making is like a traffic light.

with x yards to go (the bottom x-axis) at the x yardline (the top x-axis), then SF has an estimated win probability of y (the y-axis) if they go for it (the red line) or if they kick a field goal (the blue line). At the yards to go where the estimated win probability gain by going for it is positive (where the red line is on top of the blue line, 4 yards to go or fewer), Burke says SF should go for it. Otherwise, SF should kick a field goal. Thus for Burke, as for Baldwin, the strength of a decision is based on the estimated win probability gain by making that decision. But, as before, a decision with a high estimated win probability gain is not necessarily the best decision with certainty. To illuminate this point, we create a fourth down look-ahead chart in Figure 9b which looks ahead at the bootstrap percentage of a decision. For four to eight yards to go, the bootstrap percentage of either decision is less than 60%, which indicates that we don't have enough data to know which decision is better.



(a) (b)

Figure 9: Look-ahead charts for example play 3. Figure (a): Burke's fourth down look-ahead chart, which looks ahead at win probability for a potential upcoming fourth down. Figure (b): Our additional fourth down look-ahead chart, which looks ahead at bootstrap confidence.

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

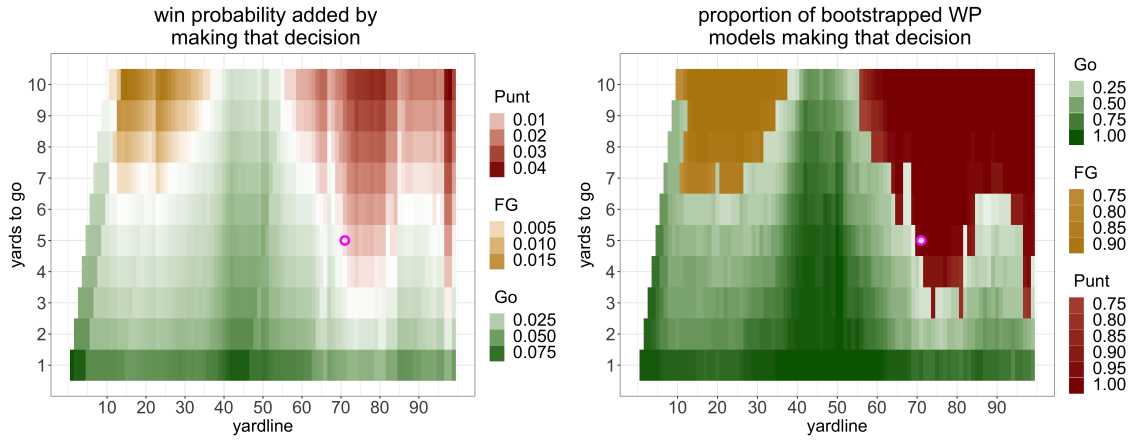


Figure 10: Our decision making for example play 4.

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

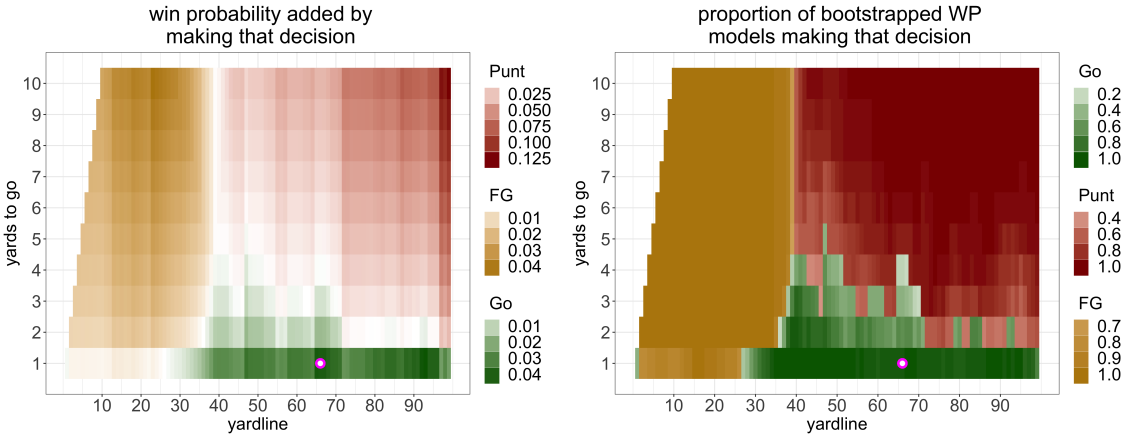


Figure 11: Our decision making for example play 5.

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

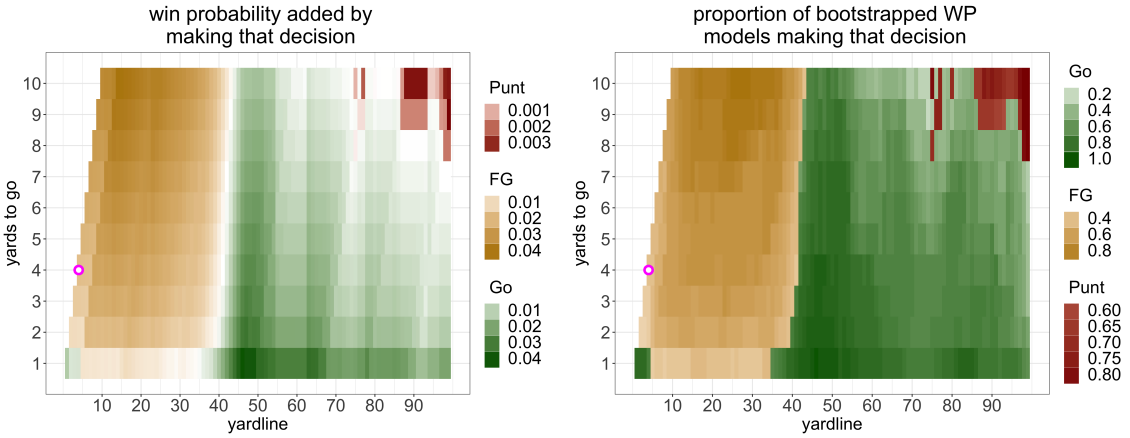


Figure 12: Our decision making for example play 6.

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

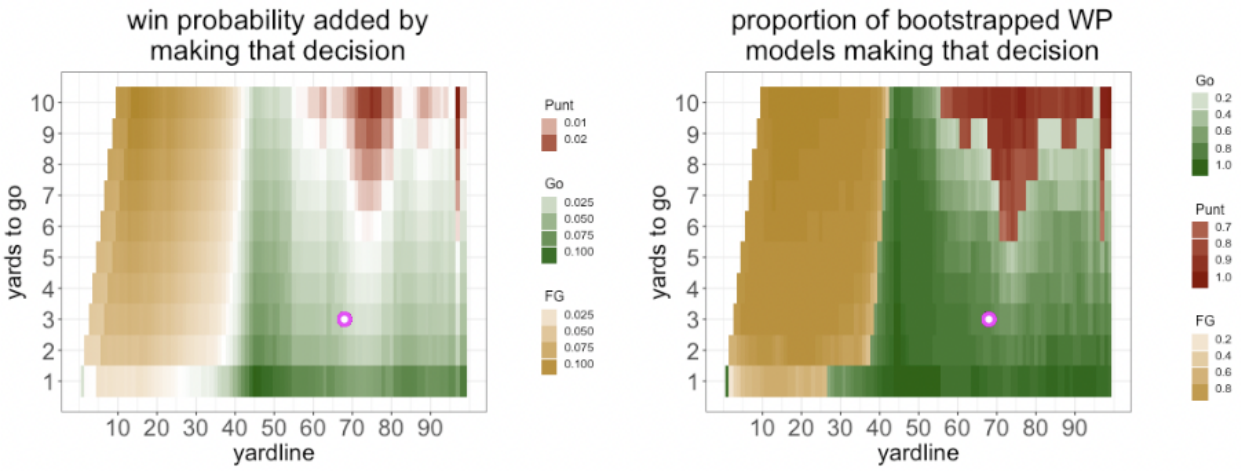


Figure 13: Our decision making for example play 7.