Analyzing 2022 NFL 4th Down Success by Team

Jack Miller and Alex Jackson

Introduction

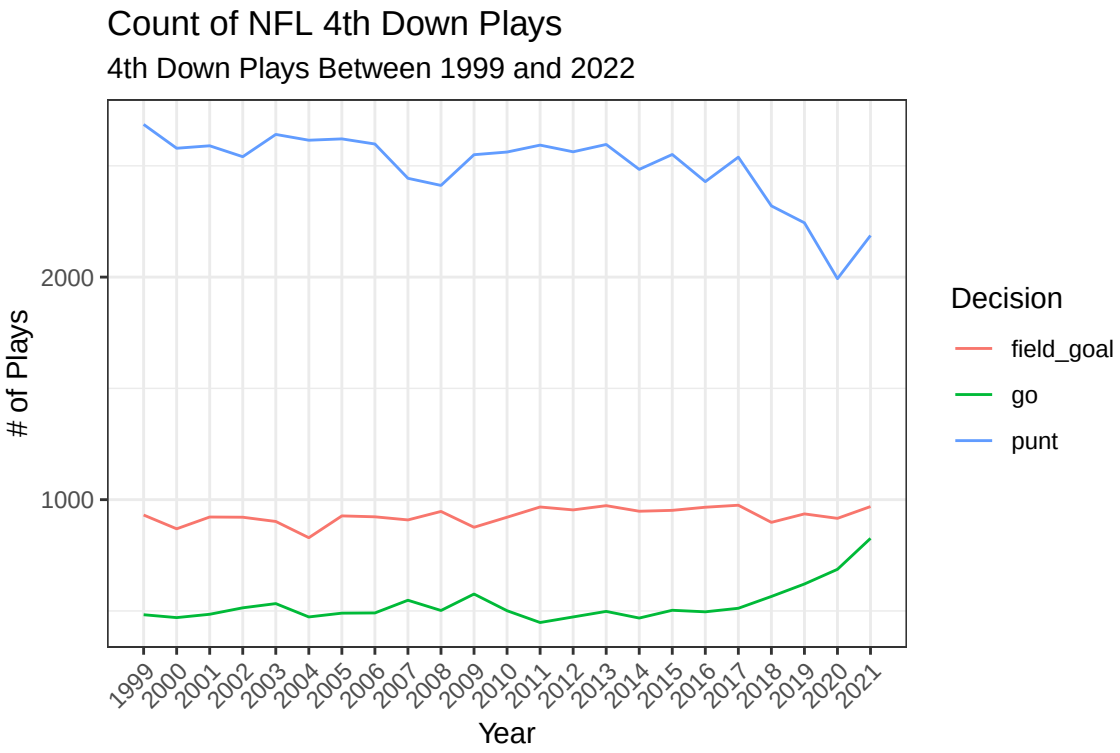
Teams across the National Football League have been incorporating analytics into their decision making more and more over the past 20 years. The emergence of analytics has been increasingly dragged by those who think football and numbers should stay separate. Recently, the decision to go for 2 points after scoring a touchdown to go down 6 instead of 7 points late in the game has become a controversy, yet it is an analytically-correct move. We believe the next enhancement of the game in regards to the incorporation of analytics comes one 4th-down decision making. Our project examines NFL 4th down decision making using play-by-play data dating back to 1999. Our goal is to build three models that predict expected win probability added for each of the three decisions coaches have on fourth downs: kick a field goal, punt it, or go for it. Personally, we believe teams should be going for it more and punting less than how they operate currently, but we wish ti support this claim with models.

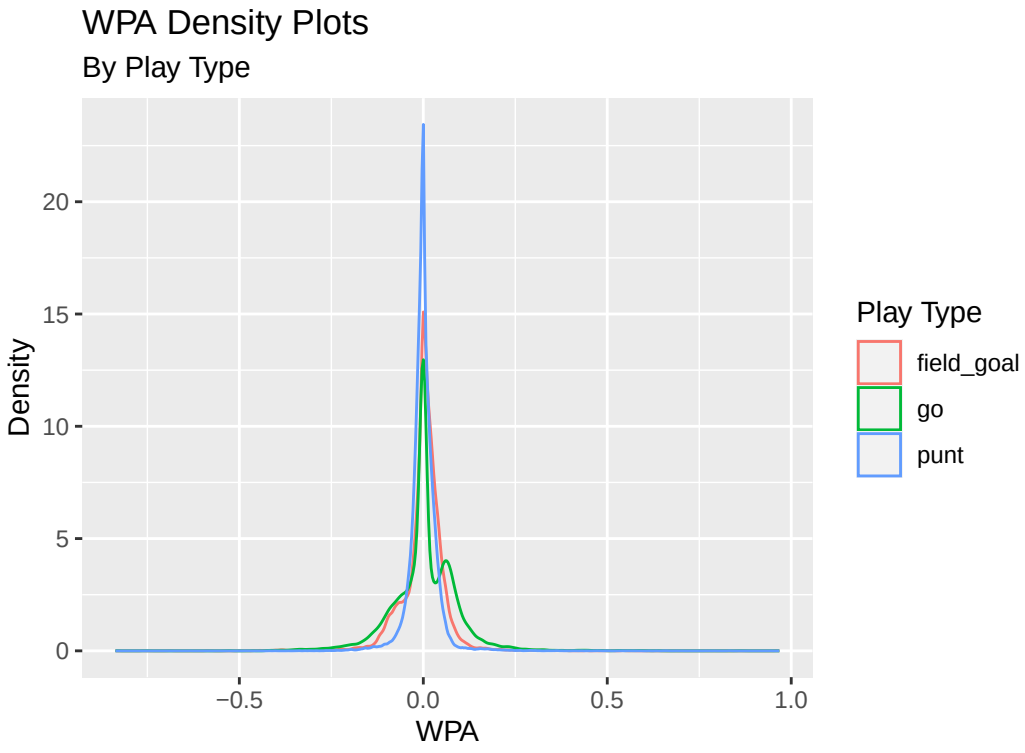
For the purpose of our study, we simplified 4th down decision making, meaning we did not differentiate between run or pass plays, and any trick plays were included as going for it (i.e. lining up in punt formation does not classify it as a punt). To train our data, we used everything that was available to us. Our data comes from the `nflfastR` package which contains data on every play dating back to 1999. After filtering for 4th down plays, we used everything from this season all the way back to 1999. Although we considered basing our project on more recent data, we figured the more data the better and decided to use it all.

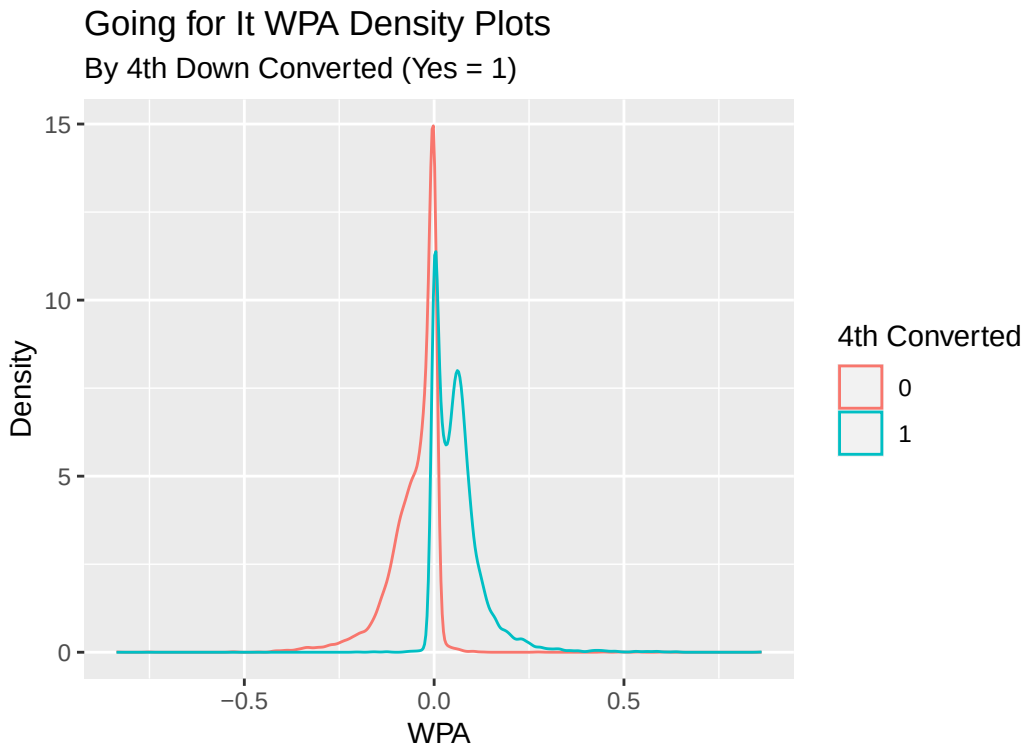
Exploratory Data Analysis

To get a better idea of what our data looks like, we first wanted to see how much data of each type of play we have. Most 4th down plays are punts, of course, and only a small piece of our dataset is go for it plays, which also makes sense. Looking at the graph, however, we can see an uptick in go attempts and a steady decrease in punts over the last 4 full seasons. This hints at the greater picture we are trying to discover: coaches haven't found the right go/punt/kick balance. But they are trending in the right direction.

play_type	count	mean_wpa
field_goal	22075	-0.00035
go	12713	0.00393
punt	59018	-0.00043


Count of NFL 4th Down Plays
4th Down Plays Between 1999 and 2022
# of Plays
2000
1000
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year
Decision
field_goal
go
punt


WPA Density Plots
By Play Type
Density
20
15
10
5
0
−0.5
0.0
0.5
1.0
WPA
Play Type
field_goal
go
punt



We also wanted to take a look at the distribution of our response variable, `wpa`. Looking at the density plots above, we can see that while the density plots of punt and kick plays look relatively normal, the density plot for going for it on 4th down appears slightly bimodal. After looking at the second plot, we can see this is due to a fourth down conversion having the biggest impact on `wpa`. This is due to successful 4th down conversions usually leading to a bigger change in WPA than a 4th down failed attempt. We believe this is because the probability of conversion is lower and teams who go for it tend to have below-average win probabilities, meaning a conversion leads to a big increase while most fourth down fails for a team already losing aren't as harmful. While we weren't sure if this was worthy of transforming WPA for the go for it model or all the models, it is something to beware of moving forward.

Modeling

The goal with our models is to create the best possible models for predicting WPA (win probability added) given a certain game state. To do this, we wanted to create a WPA model for each decision (punt, kick, go for it) that is trained on previous game states and resulting WPAs. We first split the data for each model into train and test sets using 70-30 splits for each play type. We began by training a baseline XGBoost model for each decision as well as a Lasso and Ridge regression model for each play type. After examining the RMSEs of the three different types of models, the XGBoost baseline model performed much better than both the ridge and lasso optimal models for all three play types. Because of this, we decided to use XGBoost as our model choice for all three play types. We played around with the parameters of each model to try and achieve the best predictive model possible. We did this by attempting to minimize our test set RMSE across different XGBoost parameters such as learning rate, number of iterations, subsample, and the max depth of each tree. After multiple tests using our testing data, we finally settled on a final punt, kick, and go for it model. The next step was to use these three models to predict the WPAs of critical 4th down plays.

Punt Model

Our optimal punt XG Boost model has a RMSE of 0.0216, which is the lowest RMSE we obtained throughout our XG Boost, Lasso, and Ridge Regression modelling process. Therefore, we used this as our punt wpa model.

Kick Model

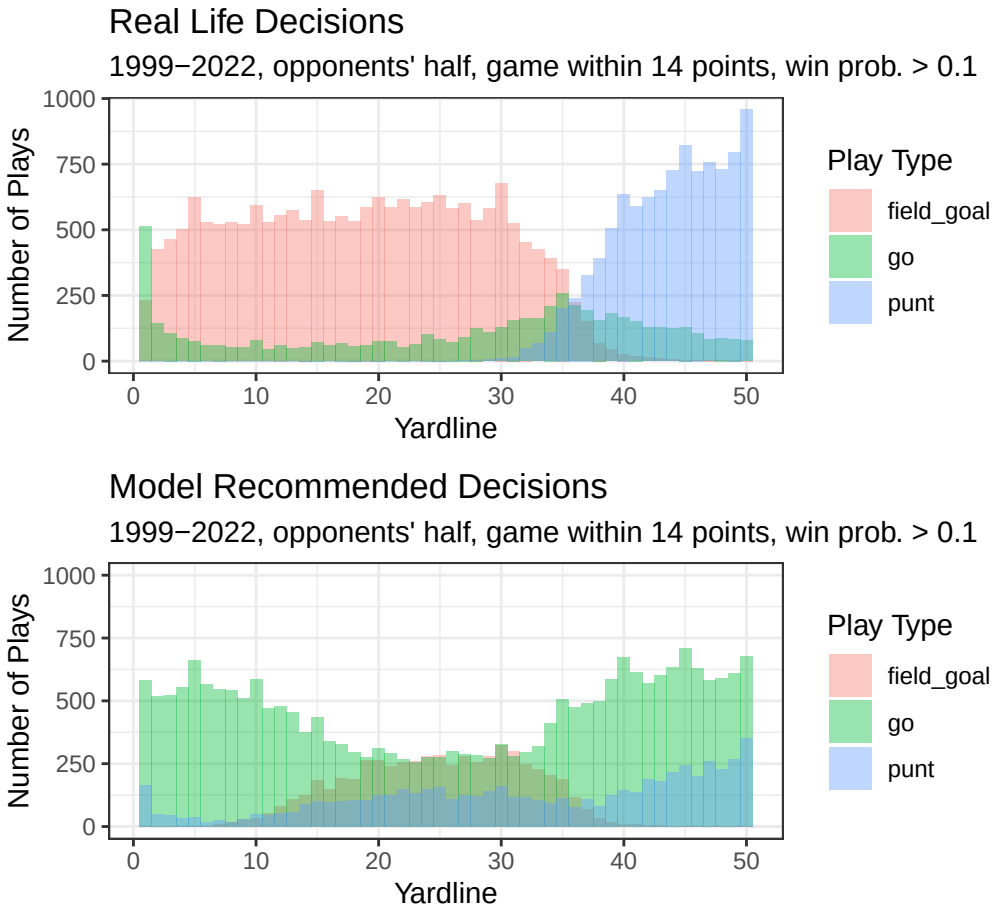
Our optimal kick XG Boost model has a RMSE of 0.03921, which is the lowest RMSE we obtained throughout our XG Boost, Lasso, and Ridge Regression modelling process. Therefore, we used this as our kick wpa model.

Go Model

Our optimal go for it XG Boost model has a RMSE of 0.05492, which is the lowest RMSE we obtained throughout our XG Boost, Lasso, and Ridge Regression modelling process. Therefore, we used this as our go for it wpa model.

Results

In our analysis, we initially applied our predictions to the entire dataset. While this would have been fine for exploration purposes, we thought it would be better to filter the data we were analyzing a bit more. Many plays in the larger dataset are "obvious" football situations where we don't need a model to tell us whether to put, kick or go for it. To get at the more nuanced plays, we created a "crucial play" dataset. This only includes plays on the opponents side of the field or within 5 yards of the line to gain, where the game is within 14 points, and where the possessing team has greater than a 0.1% chance to win. We figured this would increase the strength of our predictions as a whole and weed out some of the more pointless plays. However, we did train our models on the entire dataset in hopes that they would learn more about what it means to punt it tough situations or go for it in long to gain scenarios.

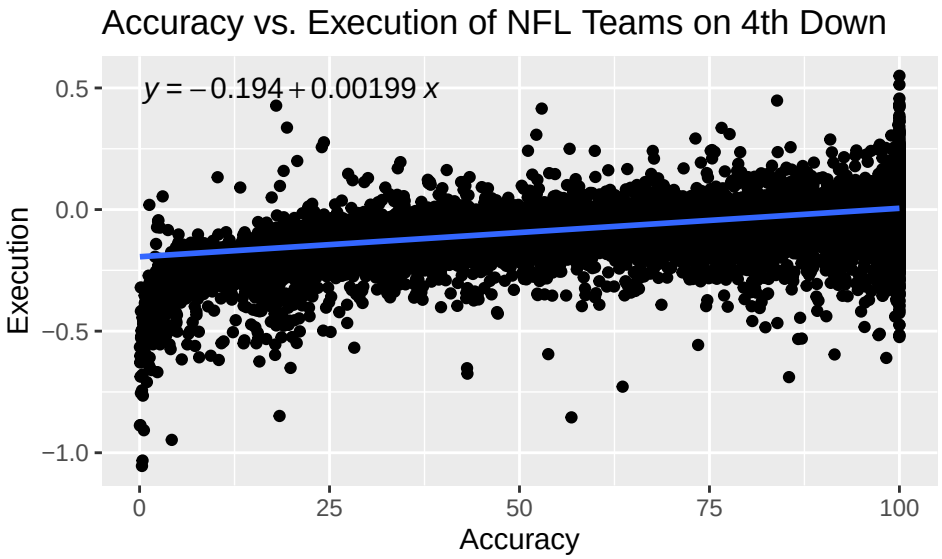


To visualize our recommendations, we plotted the frequency of each play on each yard line beyond the 50 on crucial plays. What is immediately apparent is the number of go plays. Our models clearly favors go

plays over the other types of plays, which makes sense since going for it has the most direct path to points. What's also noticeable is the extension of punt plays across the entire length of the field. The punts peak as we get closer to the 50, which is what we expect, but as seen from the plot of how the crucial plays were actually made (top) there were no punts from within the 30 yard line. Our model doesn't understand football conceptually and thus draws this big limitation. On a positive note, our model does a nice job of creating a bi-modal structure of go plays with field goals punting in the middle. This is exactly what we were hoping to see, since going for it becomes more valuable closer to the goal line and kicking a field goal becomes harder closer to the 50 yard line. The actual plays have a similar shape, with go plays peaking at the 1 yard line and around the 36 yard line, and field goals taking up a much larger portion of the area in between.

	field_goal	go	punt
field_goal	5690	11264	2375
go	127	5225	1131
punt	106	16755	4215

Our confusion matrix above, which shows the play type in the rows and our recommendation in the columns, also paints a similar picture as the graphs above. While going for it was called much less in comparison to punting or kicking, it was easily the most predicted play. We did a good job of predicting field goals and an ok job of punts, but we can see that our biggest issue was predicting to go for it too much.



Our final plot above shows our accuracy vs. execution metrics plotted against each other. The positive slope indicates that teams perform better as their play call becomes closer to our recommendation in wpa. While this is a good sign for our model as it means teams do better on average the more they agree with our model, there are other outside factors that could be playing into this.

Top 5 Executed 4th Downs

ewpa_punt	ewpa_kick	ewpa_go	recommendation	play_type	wpa	execution	accuracy
-0.138	-0.122	-0.018	go	go	0.532	0.550	100.000
0.016	-0.054	0.000	punt	punt	0.529	0.514	100.000
-0.109	-0.297	0.025	go	go	0.481	0.457	100.000
0.002	-0.172	-0.095	punt	go	0.450	0.448	83.898
-0.075	-0.424	-0.016	go	go	0.418	0.434	100.000

Above, we can see an example of the top 5 executed plays on crucial 4th downs in the last 22 seasons. Most of the plays with the largest WPA values are teams going for it probably near the end of the game, which yields the greatest WPA when converted. Our models made the right recommendation on four out of the five best executed plays, which is either a good indication or a sign that our models maybe learned too much from the plays with the highest WPA. However, the best executed plays are often times the result of extreme variance and represent the best possible outcome for a given play call making them very hard to accurately predict.

Discussion and Conclusion

Our models came with some significant limitations. For starters, each one was only trained on that specific type of play. The models don't compare to each other and have no way of seeing alternatives, it purely just tries to estimate win probability based on one type of play. This led to extrapolation in scenarios that would never exist in the NFL (like punting within the 30 yard line, kicking an 80 yard field goal, or going for it on 4th and 15 in the first quarter). Additionally, the models incurred some bias from the WPA of going for it. Of course, going for it on 4th down has the most direct route to points scored and that will increase any teams win probability no matter the scenario. There were many factors outside of game state (like positional strengths/injuries and weather) that we did not include in our model, but could have affects on WPA and decision making. Additionally, plays during the end of games have dramatic shifts in win probability that skew how our models think.

While our models and process did not turn out exactly how we wanted it to, we still ended up with a working product and learned a lot while making it. This type of modeling has a lot of potential for future application. XGBoost models can provide a predictive experience to football that with the right parameters and variables, could change the way NFL teams operate on 4th down.