Introduction

This project is looking at the current views surrounding data analysis regarding American Football and attempting to view them in a more accurate light. In recent years, NFL teams have started taking hard stances and decisive actions based on larger data beliefs. However, some teams tend to stick too hard to these beliefs (such as running having little value, going for it on 4th down, etc.) and have not had the success that some would lead people to believe. This overreliance on flawed interpretations has led to multiple teams losing games and, in turn, missing the playoffs. This can have damaging effects on an organization and its staff. This report hopes to assist any coaches who have been overwhelmed by the recent wave of data analytics in football by recontextualizing both what is the most important areas for statistical analysis derived decisions and when to deploy them.

If data isn't properly understood in its analysis, the team will suffer both on the field and financially. Having a background playing Madden at the highest level gives me a unique understanding of all the different data points being used in these calculations, allowing me to better analyze the data than most. In this project, I've reviewed over 400,000 plays from 2,500 games over 10 years to attempt to verify or deny some of these current beliefs and ultimately lead to an even more winning strategy.

The Data

The data was collected from the most recent NFL Big Data Bowl competition. After reviewing all the 10 or so CSV files, I primarily focused on the CSV file containing 255 features of play data for the aforementioned over 400,000 plays. These features contained a wide range of details that can potentially be very helpful for determining game winners. Features such as run counts, pass counts, 3rd and 4th down conversion counts, direction of plays run, and so much more. However, the data did not include certain strategic variables that would be crucial for further analysis. These features would have provided even more insight into winning strategies; however, they were not available.

winner_play	shotgun	no_huddle	qb_dropback	qb_scramble	air_yards	yards_after_catch	first_down_rush	first_down_pass	third_down	third_down
0	40	14	50	1	436	136	1	19	4	10
1	16	0	34	0	266	67	2	13	4	9
0	17	0	38	0	321	126	3	15	6	8
1	21	0	33	0	159	88	4	11	4	7
0	21	43	49	7	434	126	11	20	10	7
1	22	2	27	0	163	82	1	8	2	8
0	24	5	40	1	216	41	4	8	3	13
1	26	2	33	2	141	45	9	7	5	8
0	24	0	38	1	244	158	4	12	5	10
1	30	0	32	1	193	146	3	7	3	9

Data Cleaning

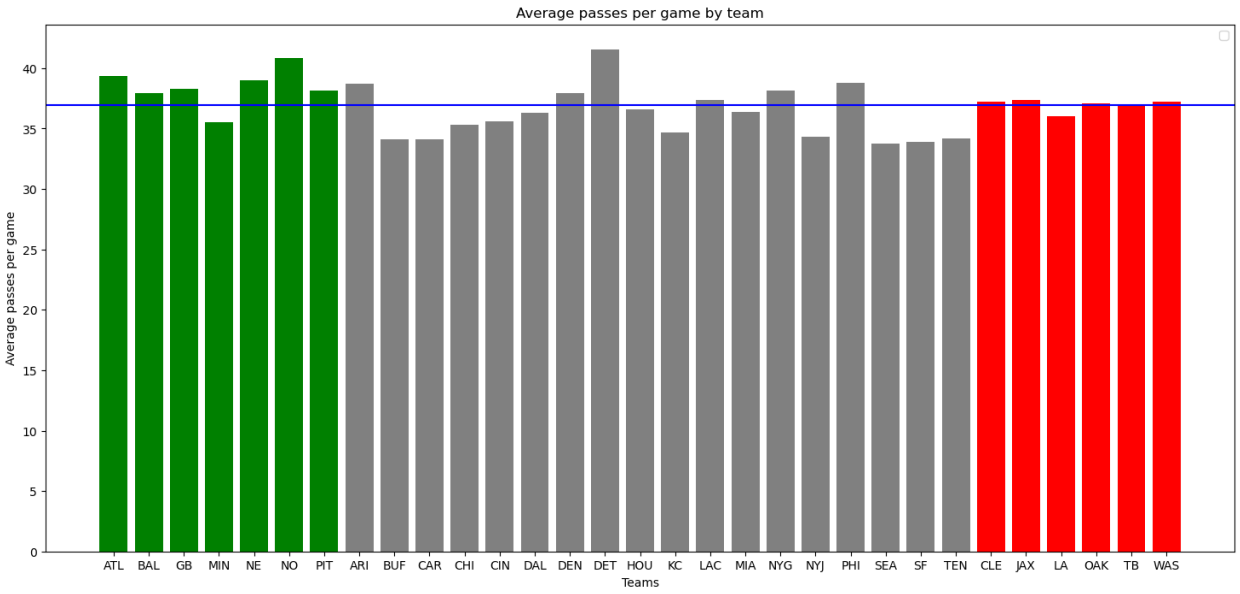
The beginning stages of the process consisted of combing through the original CSV file and removing features guaranteed to never be needed when converting the data into a data frame. Features such as scoring player name, EPA (Expected Points Added) - an already derived advanced statistic, and penalty statistics were removed. At this point, the data was brought into a Python data frame for further cleaning. From here, all null values were removed, and all plays that were not passes, runs, or field goals were also removed. Following that, it was important to drop the columns that have too much predictive power and are not in a coach's control as decision-makers during a game. These are stats like touchdowns, which are not something a coach can just decide to have more of, or turnovers, which require a mistake from the other team. It was important to just look at features like runs and passes

because a coach has access to that any time he's in a situation to call a play. Now the data is ready for modeling. The final step was to create features in order to track which plays were from the winning and losing team. This process consisted of creating columns to indicate whether a play was run by the winning team and then grouping together the total stats for each winning and losing team per game. Once the data was organized by winners and losers game stats, it was much more efficient to test with modeling how predictive these features are.

The end result is a greatly condensed data set without losing any relevant information for the study, making it incredibly easy to run multiple highly optimized models for accurate results.

Insights

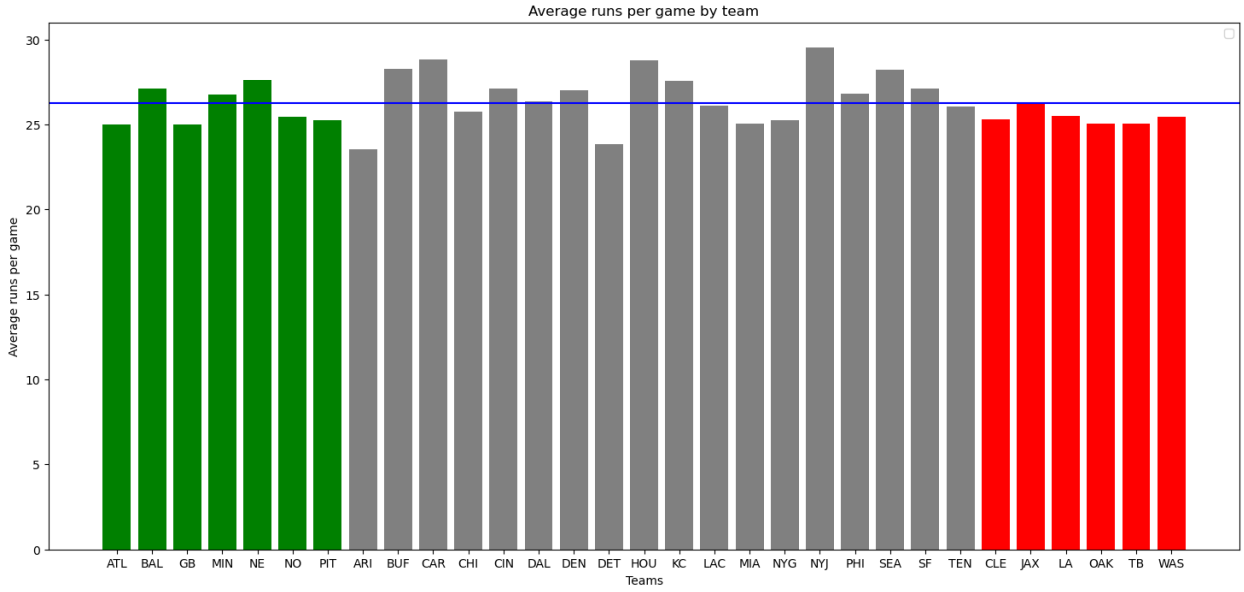
Before modeling, a quick pre-analysis was done to review the data. For the most part, nothing of any major significance stood out. Details like teams tending to pass more than they run, which also revealed was even more of a factor on 1st down, and the home teams winning more than the away teams. Once team stats started getting reviewed, something did stand out. It quickly became apparent that winning teams tended to have more pass attempts per game than losing teams.



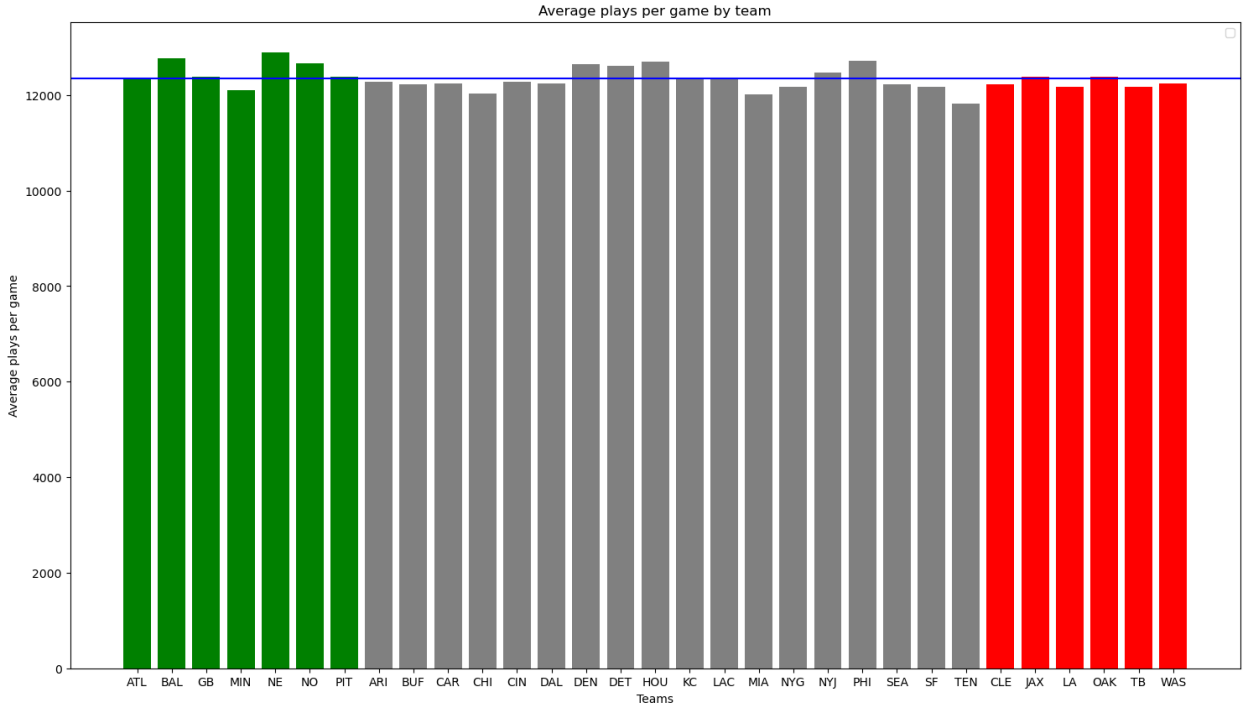
The green columns indicate the winningest teams, while the red columns are the worst preforming teams

After repeating this same analysis for runs, surprisingly the inverse was not true. To say that running was not so clearly correlated with either winning or losing. A hypothesis was formed that winning teams tend to run more plays due to them being more successful on offense, and with those extra plays, they used

more passes than runs.



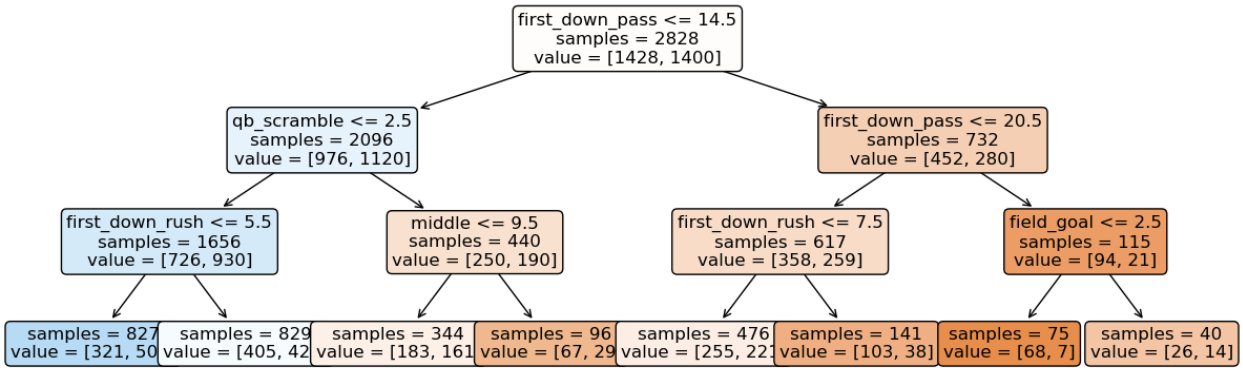
This required a final check to see the total play counts, and as it turns out, there did appear to be a connection, as we see 6 of the 7 teams with the most pass plays having above-average play counts, as shown below.



After running multiple models, the highest predictive power found from the input features was 59%. Now, keep in mind that we are measuring between 2 options: a win or a loss, meaning that our default predictive power is 50%. This leaves us with a significance from the data of about 10%, or rather, the input data had a 9% impact on the ability for the model to predict a winner or loser of an individual

game. After seeing only a 9% predictive power, a decision has been made to review the data, add back some features that may be more important, and create new ratio-based features to account for games with more or fewer plays run than average. This was done as a check to make sure nothing went unchecked and used to verify the initial findings. After going back and running those same models again with adjusted data, the strongest model prediction actually went down a percent to 58%.

All that being said, it is still worth looking at the features deemed the most impactful in the modeling process.



Here we see that the top 3 indicators of a game outcome were: 1st down pass attempts, followed by QB rush attempts, and 1st down pass attempts at a higher threshold. This corresponds well with the initial findings that teams that are successful tend to pass more and be in more first-down situations because they stay on the field with the ball. All in all, after all the data cleaning, modeling, re-cleaning, and remodeling, the models that provided a fairly low predictive power still had access to some statistics that are disproportionately indicative of success or failure. Even the strong features the models and data were able to find were more indicative of existing than how they were derived to begin with.

In short, we see that a majority of these simple input statistics do not inherently make a major difference in the outcome of the game in isolation, and ultimately doing anything very well is better than doing something with a slightly more predictive power poorly. That isn't to say that this data is so useless it shouldn't weigh into any given decision, but it shouldn't lead to making every single decision being decided by a single data point. The total scope of the project and data did not allow for further analysis at this time, but this project is one that can become much more robust over time, given additional data features and analysis. Viewing features such as coverage shells, formation, and passing concepts can be very important, especially when looking at a team-by-team basis. These can be used much more actively as opposed to passively, tracking player and coach matchups versus various strategies and using that data to exploit their personal weaknesses. Not to mention an even higher degree of predictive power for this project's overarching modeling as well.