Introduction:

Each year, hundreds of prospective NFL players are invited to participate in the National Invitational Camp, also known as the NFL Scouting Combine. According to the NFL, the purpose of the combine is to “allow NFL scouts to evaluate that year’s top draft-eligible college players on a variety of medical, mental and physical criteria.” Through the years, front offices have allowed combine results to affect their draft decisions, with the allure of drafting a "generationally talented athlete” often being too great.

The NFL began broadcasting the Scouting Combine in 2004 on NFL Network, which greatly increased public accessibility to players’ combine performances. This change not only increased fan interest in the event, but also amplified the perceived importance of measurable athletic traits among both fans and front offices. This growing attention brought to light an essential question: do raw athletic metrics captured at the combine meaningfully correlate with sustained player success in the NFL?

There are many different metrics to determine player success, with the most simple being the duration of a player’s NFL career. The average NFL career only lasts 3.3 seasons, with great variation from position to position. Three seasons is also a very important landmark for a player’s career as a player is eligible to receive pension payments once they have completed three seasons in the NFL. Though other expectations may have not been reached, a player’s career can be seen as above average if they completed three or more NFL seasons.

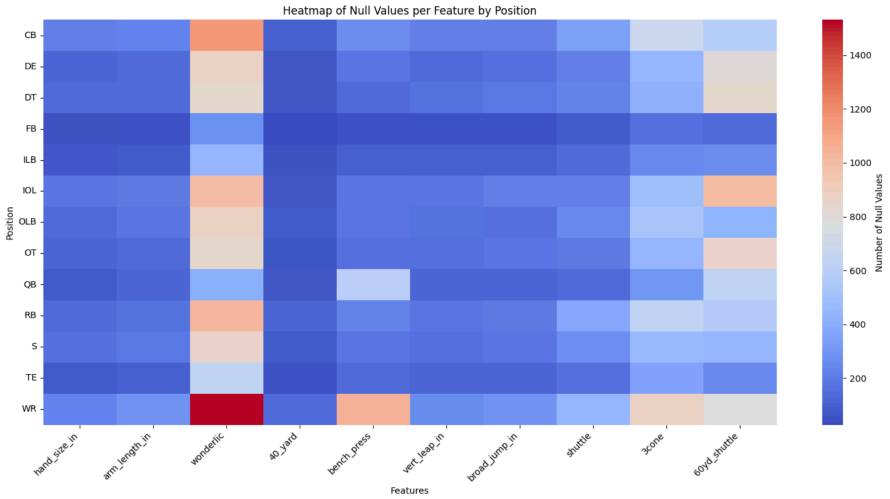
The question I will be seeking to answer is as follows: in comparing scouting combine results pre-2004 and post-2004, is there a higher correlation with a player’s career lasting longer than three years? I will be using XGBoost models on subsets of players across different positions to determine whether the increased visibility of combine results after 2004 actually translated into better predictive accuracy for career length.

Upon completing the models, it is apparent that the effect on predictive ability between pre-2004 and post-2004 models is largely dependent on the position of the players. Some positions saw little effect between the two time periods, while others saw significant change in both classes or only one class. These trends can likely be explained by changes in football playstyle over the years, and the combine results help illustrate how certain features become more or less important in predicting career length before and after 2004. Evaluation processes need to be kept more specific for each position and they will continue to advance alongside the everchanging game of football.

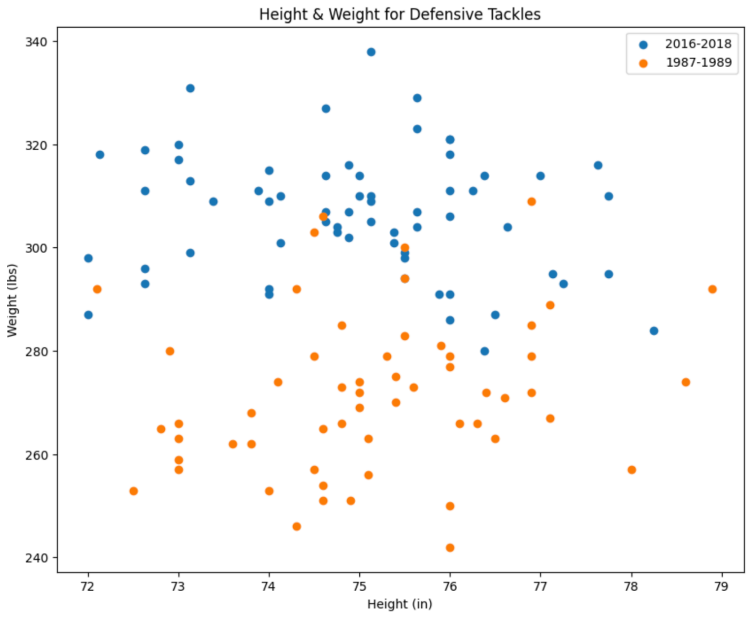
Approach:

The data containing NFL combine data was obtained from a public Kaggle dataset (https://www.kaggle.com/datasets/ulrikthygepedersen/nfl-combine-results?resource=download) which contains player combine results from 1987-2018. The following combine test results were given in this dataset: height (in), weight (lb), hand size (in), arm length (in), Wonderlic results (an IQ test often given to quarterbacks), 40 yard dash time (seconds), bench press (numbers of reps done at 225 lbs), vertical leap (in), broad jump (in), shuttle run time (seconds), 3 cone shuttle time (seconds), and 60 yard shuttle time (seconds). This data was joined with player career data provided by nfl_data_py, a Python library that allows for easy access to play-by-play, game, and player statistics.

An initial concern in using this dataset is the large amount of null values due to some players choosing not to take various tests. Their reasoning in skipping tests is often due to test results having little impact on playstyle of the position, such as vertical leap for offensive linemen or bench press for quarterbacks, but it can also be due to injuries or a strategic decision to hide a player's weakness. To address this issue, subsets of data separating players by position were used due to there being connections to positions and missing test results. The decision was also made to drop Wonderlic scores from all positions other than quarterback due to extremely high null values. This also was influential in selecting XGBoost as it works better with null values than other models.



Another important decision made was choosing to separate data from pre-2004 and post-2004. While this provided a meaningful cutoff in public perception of combine results, it also helped address a crucial concern: NFL players have grown much larger, faster, and generally more athletic over the years. A



simple way to show this is in comparing the height and weight of defensive tackles drafted between 1987-1989 and 2016-2018. Defensive tackles in the 80s had an average weight of 271.6 lbs while modern-day defensive tackles average at 306.5 lbs. If one were to try to predict an older player's metrics to a modern player, they would likely vastly underperform. While splitting the data at 2004 does not completely solve this problem as player growth happened much more gradually over the years, it does help ease this problem. Future work could be done to identify more trends in player metrics growth over the years, though it falls outside the scope of this project.

The positions I chose to focus on are quarterbacks, wide receivers, offensive tackles, defensive tackles, and cornerbacks. These positions are among the highest paying in the NFL (https://thewire.signingdaysports.com/articles/top-5-most-profitable-positions-in-the-nfl/) , making their evaluation processes crucial to team success. Quarterbacks have a very unique evaluation process from other positions and is often considered the most important player on the team. I also wanted to have variety in "skilled-positions" vs. linemen models, so I chose cornerbacks and defensive tackles for defense, with their offensive counterparts being wide receivers and offensive tackles.

Finally, I chose XGBoost as the model to be used in my analysis. XGBoost is a gradient boosting model which uses sequential decision trees, with each tree learning from the errors of previous trees. It was selected due to its strength in handling null values, which is either not possible for some distance-based models or more difficult with other tree-based models.

Each positional dataset was split into a training set (70%) and a test set (30%). The dataset contained a class imbalance in the outcome variable, with 6227 players playing less than three seasons while 5007 playing three or more seasons. To deal with this class imbalance, I chose to stratify the training and test sets. This ensures that each training/test set has an even distribution of players who played more than three seasons,

To compare my models, I will be using precision, recall, and F-1 score. Due to the class imbalance in my dataset, I will be avoiding accuracy. Another key metric used is the feature importance, which is calculated in XGBoost by measuring how much each feature improves model performance when it is used to split data. Each time a feature is used in a decision tree, XGBoost calculates and records the "gain", which is the reduction in error from that split. It then averages these gain scores across all trees, producing a score showing which features contributed most to decreasing error. This will be very important in determining which of the combine test results can be used to predict career success and how each test's importance may have changed over the years.

Analysis and Results:

Below are analyses on each position model explored. Predictive metric tables are included in appendix on page 7.

Quarterback Models:

Quarterbacks have the most unique evaluation processes, which likely lead to their model results being most unique. Precision, recall, and F-1 score all significantly improved in the post-2004 models, meaning we were able to much more accurately predict a player's longevity by NFL combine stats. In the pre-2004 model, the most important metric was calculated to be hand size, with Wonderlic score close behind it. This changed drastically in the post-2004 model, as Wonderlic score nearly doubled the next closest metric (height) in importance.

This increase in Wonderlic importance could be explained by changes in gameplay. Offensive and defensive schemes have grown much more complicated over time, with a much higher mental stress being placed on quarterbacks. They are now expected to read defensive formations, diagnose coverages and blitzes, and audible to more ideal plays in mere seconds before the ball is snapped. Interestingly enough, the NFL chose to stop administering the Wonderlic exam in 2022 due to belief that it showed no correlation to on-field results. Many quarterbacks still choose to take the test voluntarily, though it is no longer required.

Wide Receivers:

Wide receivers saw little predictive change from pre-2004 to post-2004. Precision and recall stay fairly similar across both models in predicting players to play less than three years, while there was a slight decline in predicting long career players. In both models, the 40 yard dash was by far the most important feature. This makes a lot of sense, as wide receivers skyrocket/plummet their draft stock each year based on their 40 yard dash time. Hand size and arm length grew much more important in the post-2004 model, which both contribute to what is often called a player's "catch radius". An increased catch radius means a player is able to adjust to catch passes in more difficult spots, which can create big plays and bailout quarterbacks for inerrant and inaccurate throws.

Cornerbacks:

Cornerbacks saw a rather interesting trend across the different models. The pre-2004 model was much better at predicting short careers than long careers, shown by a higher precision, recall, and F-1 score. However, this completely changed in the post-2004 model, with higher precision, recall, and F-1 score in predicting longer careers. I believe this is likely once again due to how the game of football has changed over the years. As offenses have become

more prolific in passing, defenses have had to adjust with formations and personnel groups with more defensive backs. This creates more positions for these players to play, with a need for greater player depth on each team. It would be much easier for a team to cut a cornerback when they only needed a few on the roster, but now these players are able to hang around the league a lot longer.

Offensive Tackles:

Offensive tackles saw the most drastic drop in predictive metrics from pre-2004 to post-2004, especially in predicting short career players. I believe this was likely driven from increased technicality in pass blocking for offensive tackles given the rise of game-breaking edge rushers. Offensive tackles now must possess a rare combination of size, agility, and technique. While some of these attributes can be portrayed through combine tests, others simply cannot. This is shown by decreases in feature importance for both height and weight from the pre-2004 model to the post-2004 model. The post-2004 model favors hand size and three cone shuttle speed, which can represent the mixture of size and speed needed to help slow down modern day pass rushers.

Defensive Tackles:

Defensive tackles saw a very similar trend as cornerbacks, with an increase in predictive metrics for long careers and a decrease for short careers. Similar to cornerbacks, defensive tackle is a position that many teams try to have a lot of depth at. This is in part due to the weaker stamina exhibited by many players, but is also situational as some tackles perform better in situations where the offense is more likely to rush or pass. This means that many defensive tackles who wouldn't have made it in the league have more opportunities to carve out a niche for a team.

Another interesting trend is seen in the importance of weight for defensive tackles. In the pre-2004 model, weight was the third most important feature, interestingly enough behind vertical leap and broad jump. In the post-2004 model, weight dropped to the sixth most important feature behind shuttle run time, hand size, 40 yard dash time, three cone shuttle speed, and vertical leap. I believe there is also a form of survivorship bias being observed. Pre-2004 defensive tackles' weight had a standard deviation of 18.93, but post-2004 had a standard deviation of 15.83. Most defensive tackles who are undersized are simply being filtered out before they reach the NFL combine, or possibly being switched to positions where they are more likely to succeed.

Conclusion:

Comparing the predictive performance between pre-2004 models to post-2004 models largely depends on the position of the player. Quarterbacks increased in all metrics from the older data to the new data, wide receivers saw a slight decline in long player careers, defensive tackles and cornerbacks both saw a predictive performance shift from predicting short player careers with greater accuracy to long careers, and offensive tackles saw a sharp decline in performance metrics across both career lengths. This shows that while all these positions play together on the same field, they are extremely different in terms of evaluation process and metrics which point to success.

One major limitation of this study was in simplifying reasons a player's career may be cut short. We assume that a player's career ending was due to their performance or athletic abilities not measuring up to other players. However, there are many causes that end a player's career outside of on-field performance, such as injuries or off-field scandals. Further analysis could be done to predict player injury or could factor in the reason for a player's career ending, but I chose to keep it as simple as possible for this project.

Another aspect of this study that could be explored more in future studies is understanding which NFL combine tests have little ability to project a player's career success. For example, the bench press was not found to be one of the most significant features in any of the models and often fell amongst the least important features. This result suggests that the bench press should be considered to be removed from the NFL combine, as it may not give insights into a player's future NFL success. They have done this in the past, specifically in removing the Wonderlic test due to beliefs that it did not correlate to on-field success, though that was not the findings of this study.

Overall, I believe these findings show how the game of football has changed over the years rather than the NFL combine changing. As playstyle has shifted towards pass-heavy, fast paced offenses, this had a great effect on the player base of the NFL. Quarterbacks are asked to make much quicker and more complex decisions, wide receivers are running routes to push the ball downfield more often, and defenders are rotating in and out of the game at a much higher rate to combat these changes. Players are being trained at a younger age to play football in this way, and NFL combine stats show this trend in many ways. Analysis has been done and more could be done to capture these trends in gameplay change over the years.

Appendix:

Quarterback Models:

Pre-2004 QBs:

	Precision	Recall	F-1 Score
0	0.62	0.66	0.64
1	0.47	0.42	0.45

Post-2004 QBs:

	Precision	Recall	F-1 Score
0	0.71	0.78	0.74
1	0.58	0.50	0.54

Offensive Tackle Models:

Pre-2004 OTs:

	Precision	Recall	F-1 Score
0	0.67	0.70	0.68
1	0.59	0.56	0.57

Post-2004 OTs:

	Precision	Recall	F-1 Score
0	0.47	0.46	0.47
1	0.53	0.54	0.54

Wide Receiver Models:

Pre-2004 WRs:

	Precision	Recall	F-1 Score
0	0.70	0.65	0.67
1	0.47	0.53	0.50

Post-2004 WRs:

	Precision	Recall	F-1 Score

0	0.62	0.69	0.65
1	0.49	0.42	0.45

Defensive Tackle Models:

Pre-2004 DTs:

	Precision	Recall	F-1 Score
0	0.65	0.71	0.68
1	0.55	0.49	0.52

Post-2004 DTs:

	Precision	Recall	F-1 Score
0	0.55	0.48	0.51
1	0.56	0.62	0.58

Cornerback Models:

Pre-2004 CBs:

	Precision	Recall	F-1 Score
0	0.58	0.63	0.60
1	0.48	0.43	0.45

Post-2004 CBs:

	Precision	Recall	F-1 Score
0	0.54	0.42	0.47
1	0.55	0.66	0.60