More Than The Sum of Their Parts: Evaluating Defensive Line Construction and Deployment

Krish Shah
University of Pennsylvania

Ezra Troy
University of Pennsylvania

Daniel Coale
University of Pennsylvania

Sam Pasco
University of Pennsylvania

Caleb Cannon
University of Pennsylvania

April 2023

Abstract

A productive defensive line is a key part of a successful NFL (National Football League) team. With salary cap and roster constraints, acquiring and deploying players to create a productive defensive line is a difficult task of significant interest to NFL front offices. While most attempts to value defensive line production focus on individual attribution, we propose valuing a defensive line as an entire unit, aiming to capture the surplus value generated when different types of players work together on the field. After formulating this methodology, we examine potential applications in both player selection and deployment.

Introduction

Background and Problem Description

When discussing defensive line play from a front-office/analytical perspective we consider three key problems. *Player attribution* refers to the problem of assigning value for a player's contributions on a play. *Player selection* refers to the problem of selecting players for your team – subject to salary cap and roster limit constraints. *Player deployment* refers to the problem of how to optimally deploy your players based on the game situation (both in terms of personnel and scheme).

Current methodologies and literature surrounding evaluating defensive-line play generally has a player-first approach. That is, regardless of the player attribution system you use, player selection is generally done by "optimizing" for the most value subject to your various constraints, where value is simply a player-specific score from your attribution system. The argument is that if you fill out a roster with players who have high "value" (ie. contribute more on the field relative to their cost), then you will be relatively stronger than your competition (who is getting less value for the same resources).

Key Thesis and Goal

We believe that a player-first approach has a few key short-comings:

1. **It fails to adequately attribute value gained from combinations of players working together.** That is, the value of a given defensive line unit is more than the sum of the individual players, and understanding that value can give teams further relative gains than simply looking at individual value.

2. **Player attribution is an inherently hard problem to solve.** Despite a number of statistics quantifying individual play (such as sacks, pressures, etc.), attribution is a complicated problem because while value is defined on the scale of yards (or points), those only exist as a function of what happened between the two teams on a given play, making it difficult to determine what specific contributions to the defense a player made.

3. **It struggles to provide simple, actionable player deployment strategies.** Ideally, the roster would be constructed with a specific deployment strategy in mind, in order to best take advantage of individual player's strengths. However, tying the deployment and selection processes together is a difficult task, as the two are often not approached in tandem.

Thus we propose an alternative scheme-first method, where we focus on the scheme we want to play in a given game scenario, and then find the best group of players possible that provide us the flexibility to deploy these schemes. Our method addresses the above concerns by…

1. …evaluating each play based on the combination of players on the field.

2. …placing less importance on player attribution schemes, making us more robust to potential errors in player attribution.

3. …giving us our optimal player deployment scheme as a result of the attribution and selection processes "for free".

Data

The main data used was SIS (Sports Information Solutions) charting data for all 2022-2023 NFL plays as part of the Syracuse Football Analytics 2023 NFL Blitz Competition. *This paper is a more detailed version of the model we presented during the competition.*

This dataset included information for each play of the season, including which defensive linemen were on the field (and what technique they lined up in) and the result of the play (including information about pressures, sacks, etc.) The dataset also included SIS's calculated Total Points for each defensive lineman for the season.

Additionally, we used open-source combine data to get further information about defensive lineman's athletic/physical abilities.

Metrics

We will establish a few key metrics we used throughout our analysis:

`PashRushPPS (Pass Rush Points Per Snap)`: Noise-adjusted per-snap pass rush effectiveness metric, computed as $\frac{\texttt{Pash Rush Total Points}}{\texttt{Pass Rush Snaps+70}}$.

`RunDefPPS (Run Defense Points Per Snap)`: Noise-adjusted per-snap run defense effectiveness metric, computed as $\frac{\texttt{Run Defense Total Points}}{\texttt{Run Defense Snaps+70}}$.

> We use a per-snap effectiveness metric in order to standardize for opportunity across players, as we don't want our results to be biased by (potentially erroneous) deployment decisions. Additionally, we add 70 snaps of "average" play as a prior in order to discount extreme performances observed over small windows and reward players who performed efficiently across a large number of snaps.

`EPAoP (Expected Points Added over Predicted)`

> We fit a random forest regression fitting information about the game state and offensive/defensive formation + personnel to predict the EPA (Expected Points Added) on a given play. EPAoP is the residual between the predicted value for the EPA and the observed EPA of the play.

Attribution Model

Model Description

1. Use K-Means Clustering to cluster individual players based on observed attributes, including physical attributes (height, weight, 40 yard dash), `PassRushPPS`, `RunDefPPS` and the proportion of snaps played in different techniques. Every defensive lineman is now assigned a cluster label.

2. Generate `EPAoP` using random forest regression as discussed above.

3. On every play classify the "line archetype" as the player-clusters on the field for that play. *For example, one line archetype might be two players from Cluster Two and two players from Cluster Four.*

4. Use a Mixture Density Network to model the EPAoP distribution as a function of game state and line archetype.

The output of this model is a function that given x, a vector encoding the game-state information and a line archetype l, outputs $p(y|x,l)$, the probability distribution for EPAoP conditioned on x and l.

Clustering

We identified five player clusters (hereby denoted C_0 through C_4):

C_0 was primarily made up of lighter, faster, interior linemen. C_1 was primarily made up of larger, slower (generally run-stopping) interior defensive linemen. C_2 was primarily made up of faster defensive ends who fit the profile of DEs in 4-3 schemes. C_4 was made up of faster, lighter players who fit the profile of OLBs in 3-4 schemes. C_3 was also made up of primarily-outside DEs/OLBs, bridging the gap between C_2 and C_4.

Attribute	C_0	C_1	C_2	C_3	C_4
Height (in)	75	76	76	76	75
Weight (lb)	294	316	266	254	255
40 Yard Dash (s)	4.98	5.13	4.72	4.72	4.68
RunDefPPS	.018	.016	.019	.013	.019
PassRushPPS	.014	.015	.026	.029	.030
% of snaps inside	97.2	99.9	36.4	20.9	10.9
% of snaps outside	2.8	.01	63.6	79.1	89.1

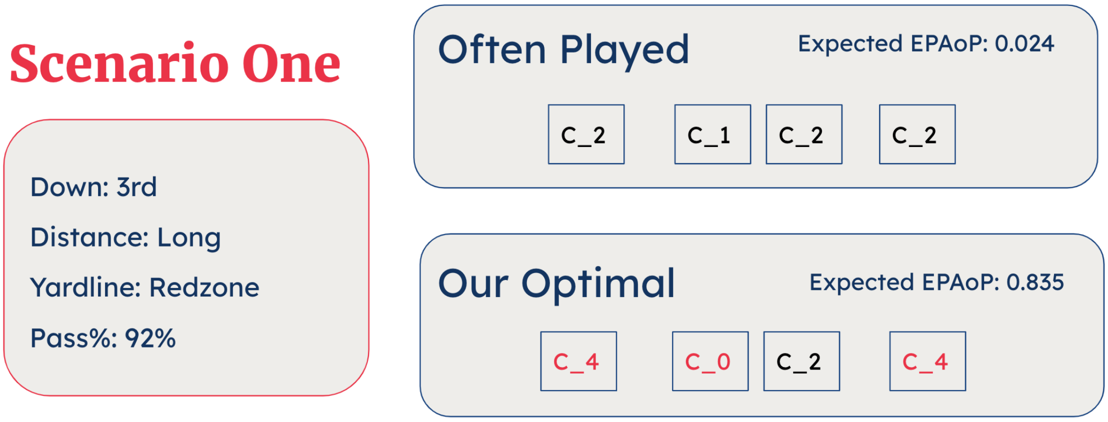
Table 1: Average Scores for Each Cluster

As clustering wasn't the main focus of the methodology, alternative clustering methods and distance functions could/should certainly be pursued, but these clusters provide a solid foundation for categorizing defensive linemen beyond simply their positional tags.

Situational Analysis

To understand the output of the model it is helpful to look at a few scenarios:

Scenario One: Get Quicker



Scenario One

Down: 3rd

Distance: Long

Yardline: Redzone

Pass%: 92%

Often Played

Expected EPAoP: 0.024

C_2 C_1 C_2 C_2

Our Optimal

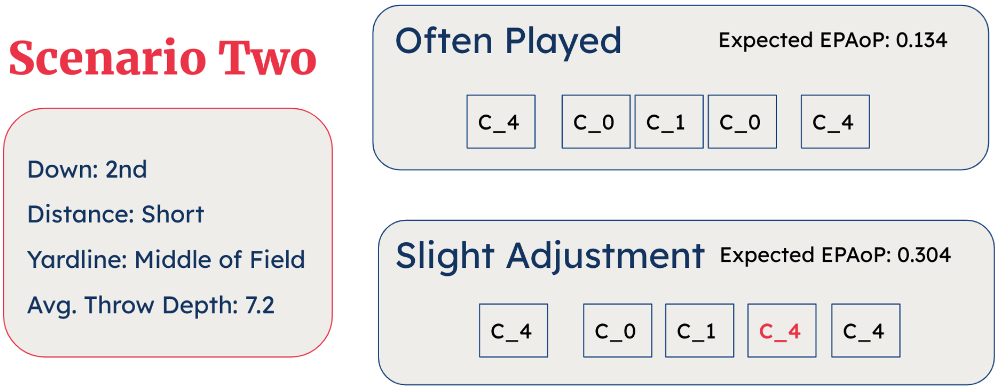
Expected EPAoP: 0.835

C_4 C_0 C_2 C_4

Third and long in the redzone is primarily a passing down, with offenses using both safe lateral passes (such as screens) and more aggressive intermediate passing plays.

Often times in this situation defenses line up in a base 4-3, playing a large interior linemen and three larger edge players. However, recognizing the passing/screen situation, our model looks for a faster interior lineman (C_0) and edge players (C_4) who can get out to the edge of the field and more effectively rush the passer.

Scenario Two: A Slight Adjustment



Scenario Two

Down: 2nd

Distance: Short

Yardline: Middle of Field

Avg. Throw Depth: 7.2

Often Played — Expected EPAoP: 0.134

C_4 C_0 C_1 C_0 C_4

Slight Adjustment — Expected EPAoP: 0.304

C_4 C_0 C_1 C_4 C_4

Second and short in the middle of the field is an interesting scenario because the "playbook is wide open." Offenses can try to move the chains running the football, or set up play-action shots to stress defenses vertically.

Often times in this situation defenses line up in a base 3-4, playing three interior linemen and two faster linebackers. However our model opts for a slight adjustment, replacing one of the interior lineman with another linebacker. This would allow a defense to better attack higher-variance play-action passes while sacrificing a little against the less-valuable (and less-variant) running plays.

Player Selection and Deployment

Model Description

Scheme Selection Model

Consider a set of game situations G. In any game situation G_i, we are able to deploy any combination of the aforementioned clusters, but you are limited to n total players. In order to ensure realistic combinations, we will require that we play at least three defensive linemen, and at least one lineman from C_0 or C_1 (the interior defensive linemen clusters). Further, for any game situation G_i, we would like to limit the probability of a terrible play (which we define as $\texttt{EPAoP} < t$) for some threshold t to be less than δ.

Our method will first select the clusters of the players we want and then choose the specific players within those clusters. Note that the clusters we select must be an n-multiset of $\{C_0, C_1, \dots\}$. Let M denote the set of all these multisets and L_i represent all valid lineup archetypes for multiset M_i.

Then for a given multiset M_j and game situation G_i:

- Define $L_{i,j} = \{l \in L_j | \Pr[\texttt{EPAoP} < t | l, G_i] < \delta\}$ (That is all lineups that meet the probability constraint). *Note that we could impose additional constraints that further restrict $L_{i,j}$.*
- Define the optimal lineup $l_{i,j} = \max_{L_{i,j}} \mathbb{E}[\texttt{EPAoP} | l, G_i]$

Additionally, we define w_i, the importance of G_i. Let S_i be the observed standard deviation of EPA on all plays in situation G_i. Then, $w_i = \frac{S_i^2}{\sum S_j^2}$. Finally we can define the optimal multiset $M_o = \max_M \sum_i w_i \cdot l_{i,j}$. When $|G|$ and $|M|$ are reasonable this can be solved naively relatively quickly.

M_o gives us the number of players from each cluster that we must choose and $l_{i,o}$ gives us the clusters we will be playing in each game situation. For a given cluster C_x let $G'_x = \{G_i \in G | C_x \in l_{i,o}\}$ (that is all the game situations this cluster will play in) and let α_i be the observed percentage of passing plays for situation G_i.

Player Selection Model

Consider a universe of players, where for each player p_i we have $p_{i,p}$ and $p_{i,r}$ (their `PassRushPPS` and `RunDefPPS` scores respectively) as well as $p_{i,c}$ (their cluster).

For every player p_j we define $p_{j,s} = \sum_i w_i(\alpha_i p_{j,p} + (1 - \alpha_i)p_{j,r})\mathbb{1}(G_i \in G'_{p_{j,c}})$. That is in every situation that this player would play, we take a weighted average of their pass rush and run defense ability (and then take the total weighted by the situational importance). We want to find the group of n players such that we match the clusters in M_o and maximise the total p_s (subject to any additional roster and salary cap constraints).

We can express this as a linear program and solve for the optimal players using any LP technique. Let x_j indicate whether we select player p_j or not. Let $p_{j,\$}$ be the cost of that player and $p_{j,b}$ some indicator (representing some other roster condition). We will use D and B to represent the maximum value of the salary and roster constraints. Then our optimization problem is as follows (we can add additional constraints as necessary):

$$\begin{aligned} \text{maximize } & \sum_j p_{j,s} \cdot x_j \\ \text{subject to } & \sum_j x_j = n \\ & \sum_j p_{j,\$} \cdot x_j \leq D \\ & \sum_j p_{j,b} \cdot x_j \leq B \\ & x_j \in \{0,1\}, \quad \forall j \end{aligned}$$

Player Deployment Model

In game situation G_i we know our optimal scheme is $l_{i,o}$. Recall that $l_{i,o}$ will give us a set of clusters to deploy. If there is no choice among players (eg. $l_{i,o}$ requires us to play two players from C_2 and we only have two players from C_2) this is simple. When we have a surplus, play players in decreasing order of their situational score: $\alpha_i p_{j,p} + (1 - \alpha_i) p_{j,r}$.

Example

Consider a toy example where $n = 6, t = -0.5, \delta = 0.1$. We will consider all NFL defensive linemen for the 2022-23 season. Assume that every NFL player is paid their 2022-23 cap hit, there is a $\$40,000,000$ budget and there can be at most three rookie contracts (to mimic a realistic NFL defensive line).

We will consider 81-unique game situations, using combinations of down, distance, yardline, and score buckets:

- Down: 1st, 2nd, 3rd/4th
- Distance: $x \leq 3$ (short), $4 \leq x \leq 8$ (medium), $x \geq 9$ (long)
- Yardline (denoted as distance to opposing goaline): $x \leq 20$ (redzone), $21 \leq x \leq 84$ (middle of the field), $x \geq 85$ (backed up)
- Score: offensive leading, offense trailing, tied

The result of the scheme selection model was $\{C_0, C_0, C_3, C_4, C_4, C_4\}$. Using this, the player selection model chose the following players:

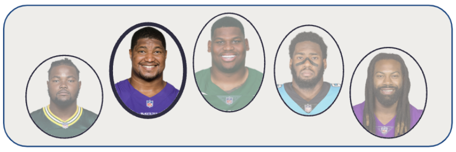
Player	Position	Cluster	Annual Value	Rookie Deal?	PassRushPPS %tile	RunDefPPS %tile
Quinnen Williams	DT	C_0	$8,132,343	Yes	93	77
Calais Campbell	DT/DE	C_0	$6,250,000	No	82	82
Brian Burns	DE/LB	C_4	$3,385,046	Yes	92	98
Dante Fowler Jr.	DE	C_4	$3,000,000	No	99	68
Za'Darius Smith	LB	C_3	$14,000,000	No	89	97
Rashan Gary	LB	C_4	$3,969,328	Yes	100	88

Table 2: Information About Chosen Players

In order to understand the situational deployment model, we can look at an example of a third down play in the middle of the field.

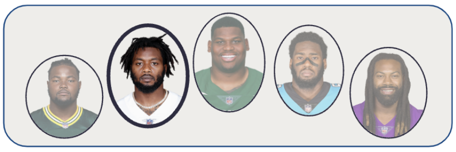
3rd and Short

Middle of Field



3rd and Long

Middle of Field



On third and short in the middle of the field, the model deploys Calais Campbell, opting to play an interior player in a short-yardage situation. On third and long in middle of the field, the model opts to play Dante Fowler Jr. instead – playing an extremely effective pass rusher in an obvious passing situation.

Model Limitations + Future Considerations

There are few natural next steps to improve the model that would make it more suitable for practical use:

The model could be improved by training the attribution model on higher fidelity data (eg. tracking data). Higher fidelity data would allow our model to gain more insight into how players interacted with one another – potentially allowing us to go into more detail schematically within the deployment model. Additionally, we used combine data to estimate the athletic/physical abilities of defensive linemen, but more accurate testing data would improve our understanding of the play of linemen who have changed physically since the combine (especially long-time veterans). We could also incorporate information like sprint speed and get off time to further differentiate players. This would allow for more precise clustering. Furthermore, having access to more seasons worth of data would increase the size of the training set and allow us to learn more sophisticated models (while decreasing the uncertainty in our original models).

Because the approach we have outlined is fairly modular, incremental improvements can be made to each step without having to rebuild the entire model. For example, more sophisticated clustering techniques could be implemented on top of the same attribution, selection and deployment models. By individually refining the techniques used at each step, the overall model will become stronger.

Lastly, we might implement opponent-adjustments into the attribution and deployment models. This would include adjusting the "value" of a defensive play based on the offensive quality it is facing (which is not unlike the adjustment that EPA already makes for game situation), as well as adjusting the scheme being deployed based on opponent tendencies/strengths.

Conclusions

We developed a scheme-first methodology for evaluating, selecting and deploying players on a defensive line. This approach provides three key benefits over traditional player-first approaches:

1. It captures the additional value generated by having certain combinations of players working together on the field at once. This not only can help teams find relative value (because they can get the same performance from potentially "worse" players) but allows teams to build defensive lines that will be more robust to individual players missing time.
2. It is more robust to differing player-attribution metrics. Because the methodology focuses on the scheme, uncertainty in player attribution will not lead to as much uncertainty in the model output as a player-first selection/deployment model.
3. It provides deployment schemes as a byproduct of the selection process. This means the model automatically considers how it would deploy players when deciding which players to select, giving it an advantage over methodologies where these processes are separate and have to be heuristically intertwined.

References

Pro-Football-Reference. (2023) *NFL Combine Results 2009-2022* [Data set]. https://www.pro-football-reference.com/draft/

NFL Combine Results. (2023) *NFL Combine Data* [Data set]. https://nflcombineresults.com/

Sports Info Solutions. (2023). [Play by play data for 2022-2023 NFL defensive lines] [Unpublished raw data]. Syracuse Football Analytics Blitz.