Machine Learning Tool for Judging Expected Longevity of NFL Defensive Linemen

Summary

The goal of this project was to create a model that could be used to determine whether a player should be invested in via a long-term contract extension. Different Web scraping techniques were utilized to gather the necessary variables/ data points that were judged as the most important when determining whether or not a players career will last at least 8 years(the target variable). After compiling/cleaning/engineering I put the variables into a logistic Regression model, which performed modestly with an accuracy score of 71%. However, after a more detailed analysis of the predictions it performed closer to 90% when accounting for players whose careers change dramatically from either injury or increased playing after their rooking contract.

Data Collection

Part 1:

First, I wanted a variable that captured athleticism. The idea being that athleticism peaks early in a player's life and tapers off as they get older. So, the higher the peak the more time a player has before they are no longer athletic enough to be in the NFL. Barring in house data from the NFL there is not a whole lot for this other than the NFL combine, which started in 1987. Luckily, nflcombineresults.com has recorded data from every player that has ever participated in the combine. To collect this data efficiently I created a loop that got the results from every Defensive Linemen that participated in the combine from 1987 to 2018. I chose 2018 to be the cutoff because I only wanted players that have had enough time to play 7+ seasons to be used in the model.

Part 2:

Now that I have the data, I wanted to create a single variable that captured a player's athleticism. Luckily, the NFL has already created this as seen here: https://www.profootballnetwork.com/what-is-ras-explaining-athletic-testing-metric/

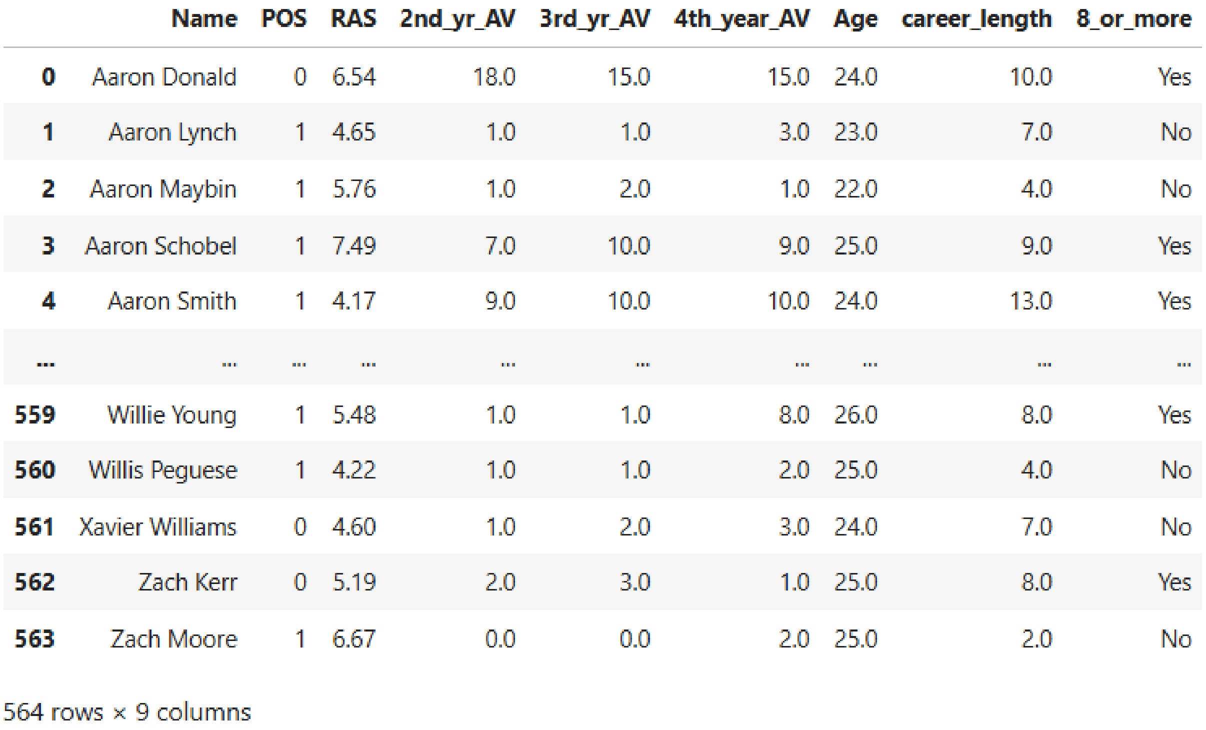
However, there was not an effective way to extract this data from either nfl.com or profootballreference.com, but there is a way to calculate this using the combine metrics. https://ras.football/ras-calculator/ has a calculator that takes in combine data and outputs RAS (Relative Athleticism Score). The formula for this score is not provided, so I created a loop that inserted each players metrics into the url of the calculator, which then sends you to a page that shows the RAS score. However, this page is composed of an image that can't be directly extracted. To counteract this I went to the page source of that url to find the unique string that came before and after the RAS score and by doing that I was able to extract the score for every player in my dataset.

Part 3:

Now, I also wanted an all-in-one variable to judge a players performance in a season. The best source for this would be pro football focus, however they have only been recording this information since 2006, so they are not an option. The next best option is pro football reference for their AV score (Approximate Value). Approximate Value is a stat that is supposed to quantify a player's performance for a team over an entire season. It is against their websites terms of service to extract this information, so instead I subscribed to their other website stathead.com. From there I just downloaded files for every defensive lineman that have ever played in the NFL and just merged it with my other data set by the players names.

Part 4:

I now have all the variables I want to make prediction, all I need is the target variable (career length). NFL.com allows for a simple web scraping method. Using this URL template 'https://www.nfl.com/players/{player-name}/' I can access every player in my dataset via a loop. For simplicity I chose to use the page source again to extract the Experience of each player. After cleaning and data engineering this was the final data set for my model.



	Name	POS	RAS	2nd_yr_AV	3rd_yr_AV	4th_year_AV	Age	career_length	8_or_more
0	Aaron Donald	0	6.54	18.0	15.0	15.0	24.0	10.0	Yes
1	Aaron Lynch	1	4.65	1.0	1.0	3.0	23.0	7.0	No
2	Aaron Maybin	1	5.76	1.0	2.0	1.0	22.0	4.0	No
3	Aaron Schobel	1	7.49	7.0	10.0	9.0	25.0	9.0	Yes
4	Aaron Smith	1	4.17	9.0	10.0	10.0	24.0	13.0	Yes
...	...	...	...	...	...	...	...	...	...
559	Willie Young	1	5.48	1.0	1.0	8.0	26.0	8.0	Yes
560	Willis Peguese	1	4.22	1.0	1.0	2.0	25.0	4.0	No
561	Xavier Williams	0	4.60	1.0	2.0	3.0	24.0	7.0	No
562	Zach Kerr	0	5.19	2.0	3.0	1.0	25.0	8.0	Yes
563	Zach Moore	1	6.67	0.0	0.0	2.0	25.0	2.0	No

564 rows × 9 columns

Variable definitions

POS: 1 = Defensive End, 0 = Defensive Tackle

RAS: Relative Athleticism Score, which is an all-in-one variable to capture a player's athleticism based on his performance at the NFL combine. Scale is from 0 to 10, 10 being the best performance possible

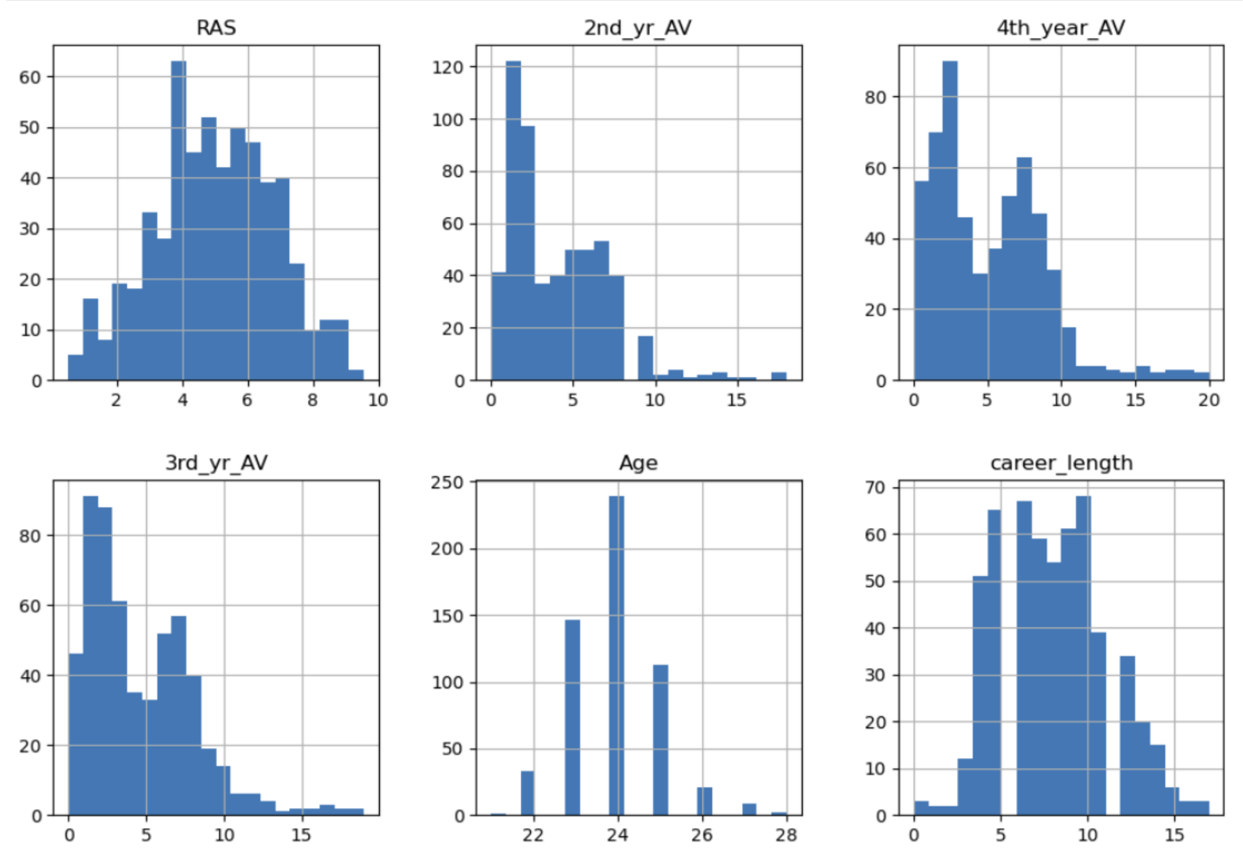
AV: Approximate value, which is an all-in-one metric used to capture a player’s performance for that season. AV scores were recorded for each players 2nd 3rd and 4th years.

Age: player’s age at the start of his second season

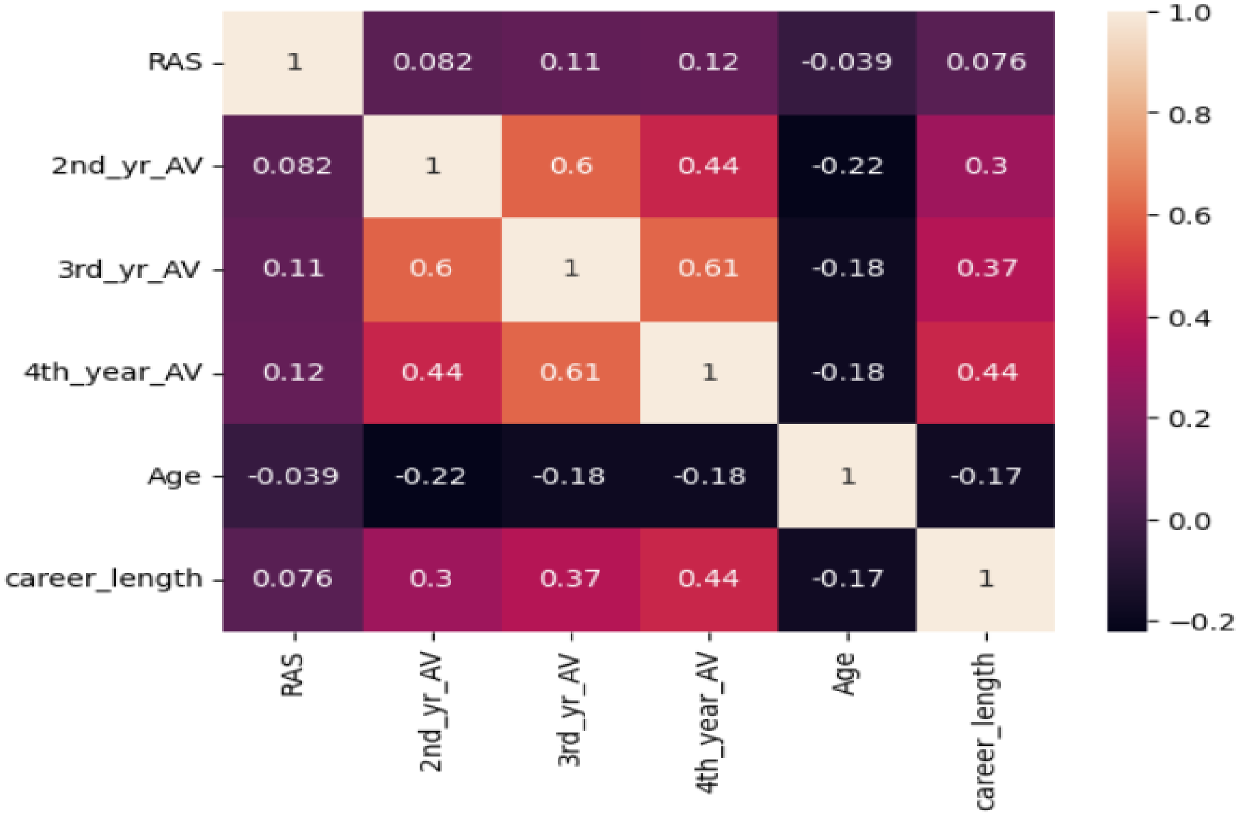
Career length: How many seasons each player played in the NFL

8_or_more: Binary variable that just states whether a player’s career lasted 8 seasons or more.

Data Exploration



Age, career_length and RAS appear to be mostly normally distributed, however the AV scores seem to be bimodal. This is because the peak to the right mostly consists of starters, while the peak to the left consists of bench/rotation players.



No surprises in this correlation matrix, players that performed well further into their rookie contract tended to have longer careers. RAS was not a huge mover which is a little surprising.

Model Selection and Results

Due to the limited variables and the dataset only have 564 samples I chose to just use a logistic regression model. After scaling and fitting/transforming the training set on to the logistic regression model I made binary predictions using the test set. Here are the metrics from that prediction.

Accuracy: 71%

	Precision	Recall	F1-score	Support
No	0.67	0.71	0.69	52
Yes	0.74	0.70	0.72	61

	Actual positive	Actual negative
Predicted positive	37	15
Predicted negative	18	43

On the surface it would seem the model did not perform that well, so I decided to research the 33 players that got misclassified to see potential reasons for their misclassification. Below is the table of that analysis.

Name	Actual	Predicted	Why?	Unpredictable
Sedrick Ellis	No	Yes	Decided to retire	1
Kenny Mixon	No	Yes	Potentially released due to off field issues	1
Greg Jefferson	No	Yes	Tore his ACL	1
Carl Simpson	No	Yes	Switched teams and fell off	0
Dave Ball	Yes	No	He never got a chance to play till years later	1
Esera Tuaolo	Yes	No	Not sure how he stayed in the league so long	0
Jason Hatcher	Yes	No	Late bloomer, potentially was sitting behind a better starter for awhile	1
Adam Carriker	No	Yes	Knee injury	1
Dennis Brown	No	Yes	model could not predict him	0
Kenyon Coleman	Yes	No	Did not get to start till after changing teams	1
Christian Covington	Yes	No	career backup	1
Luis Castillo	No	Yes	Broken leg	1
Burt Grossman	No	Yes	Neck injury	1
Darius Philon	No	Yes	Legal issues	1
Antwan Odom	No	Yes	Torn Achilles	1
Wallace Gilberry	Yes	No	Does not seem like he got a chance to start	0
Mathias Kiwanuka	Yes	No	model could not predict him	0
Jovan Haye	No	Yes	model could not predict him	0
Cam Thomas	Yes	No	career backup	0
Kevin Henry	Yes	No	Age outlier	0
Cassius Marsh	Yes	No	career backup	1
Adrian Dingle	No	Yes	model could not predict him	0
Nick Hayden	Yes	No	Late bloomer and got hurt his 4th year	1
Tony Bennett	Yes	No	Injury and holdout his 4th year	1
Nick Eason	Yes	No	career backup	1
Elvis Dumervil	Yes	No	Something happened that registered his 5th year as his 4th year and he was out that year	1

Jeff Zgonina	Yes	No	did not start till he was 30	1
John Thierry	Yes	No	low av but got a decent amount of sacks	0
Brentson Buckner	Yes	No	Traded, but slow because of injury	1
Alan Branch	Yes	No	Bigger role on new team	1
Kenny Davidson	No	Yes	productive till his final year, injury?	0
Cedric Woodard	No	Yes	Injury and holdout his 4th year	1
Timmy Jernigan	No	Yes	broken foot	1
			TOTAL SUM of unpredictable outcomes	22

This means that up 22 of the 33 misclassifications were due to extreme circumstances that are mostly unpredictable with the given data. Adjusting the results with this information in mind we end up with an accuracy score of 90/101 = 88%.

Conclusion

The model when put into the correct context performs well. The context being the player being evaluated needs to fit these 2 criteria:

1. They need to have had a high number of snaps prior to their contract negotiations.
2. They need to be mostly healthy in at least their last year on their contract.

If the corresponding player passes these two criteria, then this model will serve as a useful tool for predicting a player's longevity and whether he should get a contract extension.

How could this be improved?

1. Get more data, possibly use the entire NFL roster from 1987 to 2018 instead of using just the Defensive Linemen
2. Get per snap data for a player's performance instead of the AV score, this might counteract the model being biased towards starters and against injured players. Also, if this is done make sure to give more weight to players that play more snaps, because they would be less fresh than low snap players.
3. Find a way to replace RAS with an in-game athleticism score. This would be a better indicator of a player's functional athleticism.
4. Do a sentiment analysis of a player's technique from each player's scouting report or reduce the sample size to players from 2006 or later and use pro football focus to get the more advanced player metrics.