KNN and the NFL Combine™: Can machine learning predict offensive lineman draftability

John Barber-Ormerod, Gianluca Iarrusso

Introduction

The NFL is the highest point in professional football, each year a draft is held to add new talent to teams. The Draft consists of 7 rounds with each team having 1 selection, which they may trade away before or on draft nights. There are an additional 32 picks awarded to teams based on the players they lose in free agency. With about 250 selections each year, there is great stock placed in scouting new players to best address team needs. Traditionally this is done through watching game film and attending pro/senior days at university campuses. The combine is also an event held before the draft that is used to determine a player's ability. 6 events, as well as position-specific drills, occur for each player invited. Some players opt-out of drills to both their benefit and detriment. This paper aims to determine if the results of the 6 combine drills, as well as measurables like height and weight, can determine a player's draft value. The focus will be placed on offensive lineman(center, guard and tackle).

Literature review

40-yard dash

The individual must sprint 40 yards from a sprinter's stance. 10 and 20 yard interval times are also recorded.

20-yard shuttle

The individual runs 5 yards to his right, touches his hand to the ground, he then runs 10 yards left and touches the ground, then running right 5 yards to the return position

3-cone drill

The individual runs 5 yards and touches the ground next to the pillon, runs back to the starting pillon and touches the ground, then he runs back to the second pillon and turns 90 degrees toward a third cone, he then wraps around the cone and heads back to the first cone, again making a 90 degree turn at the second cone to return back.

225lb bench press

The individual will attempt to bench press 225 lbs as many times as possible, only reps with a full range of motion are counted.

Broad jump

The individual stands behind the start line, they then must jumbo as far as possible while still having the balance to remain in their spot. Bending of the knees and swinging of the arms for momentum are allowed.

Vertical jump

The individual stands underneath a bar with multiple swinging dowels, the jumper will stand straight up with their hand extended, the last dowel is adjusted to the persons hand. They then jump aiming to hit the highest possible dowel, the number of dowels moved correlates to the distance jumped.

KNN

KNN is a machine learning algorithm mainly focused on the classification of entries based on their proximity to other similar instances in the dataset. Unlike most other models such as linear regression or support vector machines, KNN does not require a split between a testing set and a training set. Rather it will be able to immediately classify new instances as they are added to the dataset. It does this by using a formula to measure the distance to each instance in the dataset and returns the closest K instances, it then "votes" on the classification of the unknown by taking whichever class appears most often in these K instances. The tuning process for the KNN model relies heavily on ensuring the data is properly formatted. That is to say the data must be normalized already or have similar magnitudes of values to ensure that when the calculations are done no column dominates the others by being more than three magnitudes greater than our other features. Other parameters to tune are the value for K which has a mathematical starting point of n where n is the number of instances in the dataset. Another parameter to tune is the equation used to calculate distances of which we have Euclidian, Chi Square, and Minkowski. Euclidean and Minkowski are study to perform best over datasets that have either categorical or numerical values whereas Chi Square will perform best over mixed type datasets.

Hypothesis

The hypothesis is that offensive lineman combine results can be manipulated in such a way that a KNN classifier can predict wheter or not a player could be drafted with 75% accuracy or better

METHODOLOGY

Data Gathering and Preprocessing

The values of combine information was gathered from *Pro-Football-Refrence(PFR)*. The Combine results availabale from 2000 to 2021, and contained each position group. Attaining the data was simple, *PFR* allows for each year to be downlaoded as a CSV. From there removing non-offensive lineman is trivial, there were some anomilies in our dataset. These were easily handleable, there were two instances of weights being 100lbs lighter, additonaly one height was a foot shorter. Cleaning the data was overall straight forward, the data was manageble to verrify outliers manuallyand the above issues were rectified immeditattly. Although we had access to 21 years of data, it wasdecided to onlyuse the most recent 5(2021-2016) years. This is a result of the everchnaging landscape of the *NFL* and the demands of offensive lineman. Within our data set there are instances of missing data as a result of players opting-out of the events. The decision was made to retain all participants, when testing models if the indiovidual did not have values for all features the individuals record was dropped. This was done in an effort to provide as much data as possible to the models. In total there are 330 entires with only 190 participating in all events.

Feature Selection

When examining the features we currently had, there was very little difference between the 40 yard time of a drafted and undrafted player. The idea of examining combine values and determining a player's outcome is not uncommon. An important study was by Meil in which he outlined the corelation matrix for combine events relating to offensive linemen. We created a similar matrix using our data set which is represented in figure 1. This inconjuction to events deemed important by Meil and Gallagher were the baselines for features beign fed to the KNN model. Meil placed high stock in the 40 yard dash time with minor importance placed on the shuttle, 3 cone and maximum bench press. Testing several models with features mentioned in Meil produced favorable outcomes, a model with a feature set of $\{$ 40 Yard dash, Bench Press, 3 Cone$\}$ produced 75% accuracy. likewise swapping Bench Press with Broaad jump also resulted in 75%. This wasa good sign of potential for creating an accurate model. Gallagher also favored the 40 yard dash, unlike Miel, Gallagher placed value in the Verticle Jump. This model produced an accuracy of 72% These two studies presented very good insight into what metrics were important. The 40 Yard dash was a recurring theme amongst these studies, which was the first first feature used in our model. However, unlike the previous studies we adjusted the 40 times to account for the weight of the athlete, this was done by applying the formula:

$$AdjustedWeight = \frac{PlayerWeight * 200}{40YardTime^4}$$

This produced a more accurate metric for speed, as the lighter slower players and faster heavier players would be better differentiated. The formula is not perfect but for the purposes of classification was better than weight and 40 yard time on their own. The equation is from *Pro Football Focus(PFF)*, they are one of the leading analytics for football specifically the *NFL*. Another equation used was that for the one rep max. The purpose of this was to create a tangebile metric for upper body strength. The most common equation for one rep max is as follow:

$$OneRepMax = WeightRepped * (1 + \frac{NumberOfReps}{30})$$

With all players only able to lift 225lbs this equation reduces to:

$$OneRepMax = 225 * (1 + \frac{NumberOfReps}{30})$$

The remaining features were unmodified values from

DISCUSSION

	Height	Weight	40 Yard	Vertival Jump	Bench Press	Broad Jump	3 Cone	Shuttle Run
Height	1							
Weight	0.22	1						
40 Yard	0.06	0.31	1					
Vertical Jump	-0.07	-0.29	-0.59	1				
Bench Press	-0.16	0.07	-0.32	0.21	1			
Broad Jump	0.08	-0.28	-0.60	0.67	-0.07	1		
3 Cone	0.07	0.40	0.45	-0.34	-0.11	-0.40	1	
Shuttle Run	0.13	0.41	0.55	-0.42	-0.14	-0.51	0.70	1

TABLE I
*

Figure 1. Correlation matrix for combine events and measurables (2016-2021)