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

John Barber-Ormerod, Gianluca Iarrusso

Abstract

The NFL Combine is a yearly event that allows teams to quantify the abilities of college athletes before the NFL draft. The Combine is comprised of six events that measure strength, speed and agility. It is very important for scouts to be accurate when drafting players. Knn is a classification method that decides the outcome of an entry by the classification of its nearest neighbours. The aim of this paper is to determine if a KNN model can be developed to effectively predict whether a player will be drafted or not.

Introduction

The *NFL* is the highest point in professional football. Each year, they hold a draft to add new talent to teams. The Draft comprises 7 rounds with each team having 1 selection, which they may trade away for other players and/or draft picks. 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, scouting is done through watching game film and attending pro/senior days at university campuses. The university holds pro or Senior days to display their students' skills, each institute decides when they hold one. The combine is also an event held before the draft that is used to determine a player's ability, held by the*NFL* and are 4 consecutive days. 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. We will place the focus on offensive lineman (center, guard and tackle), as previous studies leave them out because of their comparatively poor outcomes to other position groups[Meil].

Overview

40-yard dash

The individual must sprint 40 yards from a sprinter's stance. They also record times for the 10 yard and 20 yard intervals, although less common to find. This event is used to determine straight line speed and an ability to get downfield.

20-yard shuttle

This event comprises 3 cones, a start cone and 2 other cones each 5 yards away from the starting cone to the right and left, respectively. 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. This event is used to determine the speed a player changes direction and agility.

3-cone drill

Like the 20 yard shuttle, this event also has 3 cones, arranged with the starting cone, a second cone 5 yards ahead of the start, and a third cone 5 yards to the right of the 2nd cone forming an "L" shape. The individual runs 5 yards and touches the ground next to the second cone, runs back to the starting cone and touches the ground. He runs back to the second cone and turns 90 degrees toward a third cone. He then runs around the third cone and heads back to the first cone, again making a 90-degree turn at the second cone to return to the starting cone. Like the 20-yard shuttle, this event is a measure of the speed one changes direction and agility.

225lb bench press

The individual will attempt to bench press 225 lbs as many times as possible. Counting only reps with a full range of motion. This event is used to determine upper body strength.

Broad jump

The individual stands behind the start line. They then must jump 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. This event is used to determine lower body strength.

Vertical jump

The individual stands underneath a bar with multiple swinging dowels. The jumper will stand straight up with their hand extended, they adjust the last dowel to the person's hand. They then jump aiming to hit the highest possible dowel. The number of dowels moved equates to the distance jumped. Like the broad jump, the vertical jump determines lower body strength.

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 make immediate classifications of new instances as we add them to the dataset. It does this by using a formula to

measure the distance to each instance in the dataset and return 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 how the dataset is formatted and the unique challenges presented in that formatting. One parameter to tune is the value for K. Some have theorized it has a mathematical starting point of

$$K = \sqrt{n}$$

[Lali] where n is the number of instances in the dataset but this isn't always the case and has shown to cause issues in real-world data sets[Zhang]. Another parameter to tune is the equation of the distance function. There are many distance functions that can be applied to a KNN model, each with their own strengths and weaknesses.

Overfitting and Underfitting

Overfitting and underfitting are two concepts in machine learning that are essential to understanding the performance of a model. The first, overfitting, means your model has become too good at predicting training data and it will cause very poor test data accuracy. Underfitting is an even worse issue where it cannot predict either. These problems are also prevalent in KNN models and come from the K value[Z Zhong]. Larger K values will underfit the model due to too much information and a small will overfit due to lack of information[Z Zhong]. A balanced K value is vital in alleviating these effects[Z Zhong].

Minkoski

The Minkowski distance formula, written as

$$Minkowski(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

is a way to calculate the distance between two vectors X and Y. The formula itself has different names depending on the value of p, where p=1 is called the Manhattan distance and p=2 is called the Euclidean distance. This formula for determining distance is very common in KNN classifiers and is even the default measurement formula for many KNN libraries.

Feature Selection

Feature selection is an integral part of the preprocessing for classification. Features are attributes used by models to predict the outcome, some features are raw data, others normalized. The goal of feature selection is to remove entropy(degree of randomness) from our data set.[Mwadulo][Jovic] This process is not trivial, there is alot of effort placed in feature selection, many methods have been created to address this issue. The most general description for the feature selection process is reducing redundant or useless features.[Jovic] The feature that are given to the model will be the only information it sees, thus it is important for an optimium outcome it is fed the leanest and most powerrful feature set.[Mwadulo] Feature selection only deals with removing of existing features, not creating new ones.[Mwadulo]

Feature Extraction

Feature extraction is an equally important component to developing a model. Feature extraction works to create new features from existing ones.[Mwadulo] This process can be done by applying functions to the values of one or more existing features. Feature extraction has the same goal as feature selection, to reduce entropy and produce an optimal feature set.[Mwadulo] It is not uncommon for both to apply to a feature set to generate a truly optimal feature set.

LITERATURE REVIEW

Combine Values determining success

Meil noted a difference between the drafted and undrafted players regarding performance in 40 yard dash, 3 cone drill and bench press.[Meil] He also references a study conducted that found that broad jump and vertical jump when paired for 40-yard dash were a good representation of draft order.[Robbins] Although the intentions of this model was not to predict the exact order, this provided a strong baseline to test against. The key point made by Meil was that in a comparison between personal and statistical decisions, the personal is usually worse.[Meil] Aligning well with our topic as the aim of this model is to remove some of the human bias from the selection process. The metric of success is more subjective than draft status, however Meil states that for success important metics for linemen are: height, weight, bench press, broad jump, and three-cone drill.[Meil] This is an interesting feature set that will be considered when selecting our own features. Gallagher points to the combine events of both broad jump and vertical jump, and the 40 yard dash as drills displaying athleticism with no in game applications.[Gallagher] Athleticism is an important metric for drafting players, it can offset a lack of technical skills, as the player is a "Diamond in the rough". Gallagher quantified that an ideal first round linemen would jump farther than 30 inches, and a sub 5 second 40 yard dash.[Gallagher] Both of Gallagher's findings are baselines for our feature set.

Distance formulas in KNN

One of the most important metrics to consider when creating a KNN model is the distance formula that will decide the similarity of two instances in the data set. There are many aspects to consider, as each formula will bring something new to the table in terms of flexibility. For instance, in a study conducted by Li-Yu et al. about distance formulas and how they interact with a data set, there were significant differences based on the data in the dataset[Li-Yu]. With Minkowski distances and their derivatives, namely Manhattan and Euclidean, they performed well on categorical datasets with a low number of features. In contrast to this, those distances would perform better on numerical datasets with more features. Their performance in mixed datasets would drop off considerably when adding more features, causing their performance to take an almost 40% dip when moving from 6 to 7. In their observations on Chi Squared in the Li-Yu et al. study, it was found that it worked best on all datasets but would perform better than others on

datasets with mixed values[Li-yu]. We should also note that choosing a p value for Minkowski will vary by data set. It was observed that there was little difference between the two in terms of performance(Hu,2016). In other studies, it was observed that the Euclidean measure was not an effective way to measure distance(Najat et al., 2019). Data type isn't the only thing to keep in mind, however, when choosing a distance formula. Another important consideration, especially in a voting algorithm like KNN, is noise in the data. A study by Alfeilat et Al. would find that formulas like Manhattan, Hassanat, and Lorentzian performed well when there was over 90% noise in the dataset. Despite these findings, however, it is important to note that these remain recommendations, since each dataset is unique in its challenges and all are subject to the "No free lunch" theorem (Alfeilat et al.,2019). The "No Free Lunch" theorem in this case being the idea put forth by David Wolpert that states the idea that if some algorithm A is better than algorithm B on some cost function, in our case accuracy, then there are equally as many times where B is better than A[Wolpert].

HYPOTHESIS

The hypothesis is that we can manipulate offensive lineman Combine results in such a way that a KNN classifier can predict whether 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 available from 2000 to 2021, and contained each position group. Attaining the data was simple, *PFR* allows for each year to be downloaded as a CSV. From there, removing non-offensive lineman is trivial, there were some anomalies in our dataset. These were easily handleable. There were two instances of weights being 100lbs lighter, an additional one height was a foot shorter. Cleaning the data was overall straightforward, the data was manageable to verify outliers manually and we rectified immediately the above issues. Although we had access to 21 years of data, we decided to only use the most recent 6 years(2021-2016). This results from the ever changing landscape of the *NFL* and the demands of offensive lineman. Within our data set, there are instances of missing data because of players opting-out of the events. The decision was made to keep all participants, when testing models if the individual did not have values for all features the individual's record was dropped. This was done to provide as much data as possible to the models. There are 330 entries with only 190 taking part in all events. With the differences between players drafted later and undrafted players being so minute, it was important to distinguish as best as possible. This was accomplished by not normalizing the data, normalizing the data would unfavorably affect our model. The lack of normalization does disproportionately value bench press, this value is an integer, often greater than 20, Decimal values like the 40 yard dash would not see a difference greater than 2.00.

Feature Selection and Extraction

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, with events deemed important by Meil and Gallagher, were the baselines for features being 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.[Meil] 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.[Meil] likewise, swapping Bench Press with Broad jump also resulted in 75%. This was a good sign of potential for creating an accurate model. Gallagher also favored the 40 yard dash. Unlike Miel, Gallagher placed value in the vertical jump.[Gallagher] 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 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 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 tangible metric for upper body strength. The most common equation for one rep max is:

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

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 the events, the 3 cone and 20 yard shuttle run. Both features had strong corelation to the 40 yard dash. There was a minor negative correlation to the bench press. However, this set was what we felt best represented what we needed for an offensive lineman. 40 yard dash showed downfield speed. The bench press was an indicator of their upper body strength to push defenders. Finally, the shuttle and 3 cone measure one's ability to change direction, important in swing blocks and reach blocks. We removed both broad and vertical jumps from the feature set, as well as the aggregated counterparts replaced 40 yard dash and bench.

	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)

Model Creation

Once these features were applied to our classifier, we had 192 entries We then needed to determine the number of neighbors applying the optimal neighbor formula:

$$OptimalNeighbor = \sqrt{NumOfEntries}$$

Resulting in an optimal neighbor number of 13. However, with some additional testing, it was determined that a K value of 17 would work better for this data set, while still not over-fitting. Other K values such as 15 and 19 both produce accuracies greater than 80% suggesting the model is properly responded to changes. It is important to ensure we are not creating a model that overfit. Using sklearn, we implemented a KNN classifier. When changing parameters, we opted to use the Mickoski Distance formula. We applied a p value of 2, essentially changing the formula to Euclidean distance, reducing the aforementioned formula to:

$$Distance = \sqrt{(\sum_{i=1}^{n} |x_i - y_i|^2)}$$

Sklearn also allows for adjusting weights, opting for the uniform weight, meaning there was no premium placed on closer nodes. Offsetting similarity between late draft picks and undrafted players.

RESULTS

The final feature set is as such:

- Weight adjusted 40 yard dash
- One rep maximum bench press
- 3 cone drill
- 20 yard shuttle run

These features and a K value of 17 Produce a model with outcomes displayed in figure 2. With the accuracy of this model being above the desired 75% we can confirm that there is a model that can predict an offensive linemen's draft status with over 75% accuracy.

	Precision	Recall	F1-Score	Support
Drafted	0.94	0.94	0.94	17
Not Drafted	0.67	0.67	0.67	3
Accuracy	90%			

TABLE II
FIGURE 2. OUTCOME OF KNN-CLASSIFIER

DISCUSSION

Although some sources could correlate linemen's ability to their performance in jumps both vertical and broad, our models were more successful in ignoring these features. The studies conducted by Meil and Gallagher were excellent starting points for this paper. Providing insight into the demands of offensive lineman as what metrics were important. Our outcome overall and for classifying Drafted players was very promising, however there is some worry as to the undrafted results. This trend extended throughout refining the model as well as existed in the models using Meil and Gallagher features. The most plausible reason for this outcome is the difference in talent. A first round player will physically perform better than a seventh round pick. However, the difference between a player taken in the 7th round and one not drafted is much smaller. It is not uncommon for seventh round picks to be cut from the starting roster prior to the start of the season, a trend common amongst all position groups. Although this model boasts high accuracy, it should be treated as another tool for scouts. All sports drafts are equal parts luck and scouting. This model may also have a very short shelf life, as the demands of players is changing. Differences between coreelations within our matrix and Meil's could indicate this change. These changes can cause a model to be obsolete quicker, as other combine features perform better for classification. There would need to be constant validation test to ensure this is the optimal feature set. It is also important to make a distinction that this model uses only combine results, as pro days are geared toward showing off the players, the combine is fair competition with times usually more accurate. This model's strength is its precision for determining drafted players. This is especially impressive with the aforementioned little difference in players' event scores. One likely reason for the miss classification is the depth of each draft. These events are random, the number of offensive lineman varies based on year, a breakdown of this is:

- 2016: 53 70% Drafted
- 2017: 47 62% Drafted
- 2018: 48 71% Drafted
- 2019: 57 60% Drafted
- 2020: 52 79% Drafted
- 2021: 73 47% Drafted

This is important to consider, with such a disparity in both % of players drafted and attendees, there is higher influence from some years compared to others. With 2021 having both the largest amount of offensive linemen attend but also the lowest draft percentage. Although combine metrics are tangible, team

needs and depth are tangible, this can be a huge factor into a player's draft status. Another potential model could be one used to determine whether a player would make the starting 53-player roster at the beginning of the season. This model would require significantly more analytics and metrics to be developed for very subjective factors, like depth at position, versatility, and team philosophy.

CONCLUSSION

Through the use of feature selection and extraction we were able to develop a set of features that when used within a KNN model where k = 17, it is 90% accurate at predicting a player's draft status. The results produced by this feature set outperformed the feature sets derived from Meil and Gallaghers studies.[Meil][Gallagher] The use of weight adjusted 40 yard dash times benefited our accuracy, likewise is true with the maximum bench press weight. Meils use of a correlation matrix translated well for this paper, creating our own gave us insight into the important features to consider and assisted in understanding potential shifts in demands for players. The lack of normalization assisted in widening the gap between drafted and undrafted players, the lack of large differences between the two groups made creating an accurate model challenging. An important result observed was that not only was the model 90% accurate but it also boasts high accuracy with different neighbours. Losing only 5% accuracy when going from 17 to 15 and 17 to 19, provided confidence that we were not overfitting the model.

BIBLIOGRAPHY

L.Y. Hu, M.W. Huang, S.W. Ke, and C.H.Tsai (2016,5) "The distance function effect on k-nearest neighbor classification for medical datasets" digitial.

N. Ali, D. Neagu, and P. Trundle (2019,11) "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets" digital. No. 1559

H.A.A. Alfeilat, A.B.A Hassanat, O. Lasassmeh, A.S. Tarawneh, M.B. Alhasanat, M.S.E. Salman, and V.B.S. Prasath (2019,12) "Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review" digital. Vol. 7 no. 4

U. Lall and A. Sharma, "A nearest neighbor bootstrap for resampling hydrologic time series," Water Resour. Res., vol. 32, no. 3, pp. 679-693, 1996.

S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang(2018,5) "Efficient kNN Classification With Different Numbers of Nearest Neighbors" digital. Vol. 29 no. 5

A.Jovic, K. Brkic, and N. Bogunovic "A review of feature selection methods with applications" digital.

M. Gallagher (2019) "A Better Predictor of NFL Success: Collegiate Performance or the NFL Draft Combine?". digital.

A.J. Meil(2018) "PREDICTING SUCCESS USING THE NFL SCOUTING COMBINE" digital.

D.W. Robbins(2011,10) "Positional Physical Characteristics of Players Drafted Into the National Football League" digital. Vol. 25 no. 10

M.W. Mwadulo "A Review on Feature Selection Methods For Classification Tasks" digital. vol.5 no.6

Z. Zhang (2016,6) "Introduction to machine learning: k-nearest neighbors" digital. vol. 4 no.11

D. Wolpert and W. Macready (1996, 3) "No Free Lunch Theorems for Search." digital.