

Data Blitz: Empowering NFL Draft Decisions through Machine Learning and Statistical Modeling

Dylan Jorling & Dareh Aslanpour

Tuesday 13th June, 2023

1 Background and Introduction

In 2021 the NFL generated over \$17 billion of revenue, dwarfing every other domestic professional sports league despite playing considerably fewer games [1]. What often gets lost from a fan’s perspective is that the league isn’t just a form of entertainment, but also a massive corporation where billions of dollars are at stake. From a team owner’s perspective, maximizing return on investment is the number one priority, while from a player’s perspective, making as much money while their body still allows them to is a primary motivation. The league is competitive and cutthroat, and trying to find an edge over the competition is done on every level.

Winning on the field generates increased fan interest, leading to more ticket and merchandise sales, larger local tv deals, and more money in owners’ pockets. Therefore, owners invest hundreds of millions of dollars each year on coaches to lead teams, scouts to properly assess talent, and players to perform on the field. Each NFL franchise has 52 roster spots and over \$220 million (and rising) to spend every year on player salaries [5]. With 32 total franchises, this means that the 1,664 players who earn roster spots will split over \$7 billion dollars in the upcoming season alone.

While the playoffs and super bowl are the highlight of the season for fans, there is no event that impacts more NFL personnel than the NFL draft. In the NFL draft, teams select from the best draft-eligible college football players in hopes of improving their current roster. Each franchise is allotted one pick per round over 7 total rounds, and the draft order is in reverse of the previous years’ standings. The NFL draft is extremely important for every team involved, and for different reasons. Teams picking towards the top of the draft performed poorly in the prior season and are often devoid of talent. While they ”earn” priority selection, enabling them to choose the most elite players, success in college football does not guarantee success in the NFL. Although top draft picks have the highest probability of being franchise-altering players, drafting a “bust”—a player who greatly underperforms his draft slot—is extremely costly and could set a team back years.

Teams picking towards the back of the draft typically have very little flexibility to improve, or even maintain, their rosters. Players on good teams often out-perform previous contracts, which makes it difficult for those teams to re-sign their players within the limit of the salary cap. These teams’ clearest avenue to improvement is through the draft, which is made even more difficult by the fact they own lower-priority draft picks. For these teams to maintain success then, they must

find “steals” in the draft who greatly outperform their draft slots.

From a player’s perspective, the difference between getting drafted or not will likely affect their entire life’s trajectory. It could be the difference between earning close to a million dollars in one year and working a minimum wage job. Getting drafted means an automatic guarantee of income. Even “Mr. Irrelevant”—the last overall pick, number 261—earns \$85k just for being drafted, and will lock in \$725k if he manages to make the final roster cut. A player drafted in the top 100 picks earns a minimum of \$940k, while a top-32 first round pick earns a whopping minimum of \$6.1M [5].

Given the draft stakes for both franchises and players, an intense pre-draft process occurs over many months prior to the actual draft. Teams are looking to identify undervalued talent while players are trying to stand out from their peers. One of the most important events in this process is the NFL combine. The NFL combine is a 4-day event where the top college football players are invited to partake in what is essentially a lengthy job interview [2]. Players are tested both physically, with measurements and drills, and mentally, with intelligence tests and long interviews. While college performance is typically the leading factor in evaluating a draft prospect, combine results are also thought to be an important part of the process.

In our paper we evaluate the predictive power of physical combine results in determining 1) whether a player is drafted or not and 2) what pick a player is likely to be drafted with. To answer the first question, we set up a binary classification problem and build a variety of predictive models. To answer the second question, we isolate only those players who are actually drafted and analyze a regression problem using draft pick as a response variable.

2 Data Cleaning and Overview

Initial combine data is sourced from nflcombineresults.com and contains 12 unique combine measurements over 13,544 players and 36 years [3]. One key feature missing from the initial data is our response variable, draft pick. We retrieve draft pick data from pro-football-reference.com over the same 36 year time period and append it to the initial dataset [4]. Due to inconsistencies with names of players, colleges and positions between the two datasets, matching names up between the two sets is non-trivial. Additionally, we do not know which players who were actually drafted did not attend the combine. We match the datasets first using name, and then by using a combination of college and year to identify cases where players have slightly different names in the two datasets. We assign draft picks to 8200 players out of the 9600 who were drafted over the 36 year period, and while we likely incorrectly assign some players, it is plausible that a majority of this discrepancy stems from drafted players who did not attend the combine.

Next, we trim positional categories from 25 down to 9, including completely dropping the specialist positions of kickers, punters and long snappers. We also remove the Wonderlic and 60 yard shuttle features since more than 2/3rds of the data is missing for each. Many of the remaining variables still contain significant amounts of missing data, and instead of trying to impute the data we decide to only keep the complete cases. The final dataset includes 5,843 observations with 10 numerical measurements each, plus the categorical position variable which we one-hot encode.

For the classification problem we assign a binary response variable based on whether a player was drafted or not. We analyze the regression problem on the 3,422 observations where players were actually drafted, and use raw overall draft pick as the response variable. It should be noted that a lower draft pick corresponds to an earlier pick and theoretically a better player. Many of the numerical variables are self-explanatory, such as height, weight, hand size and arm length. The remaining, less obvious variables are:

- 40-Yard Dash- electronically timed forty yard dash time, in seconds
- Bench- amount of consecutive repetitions of a 225-pound bench press completed
- Vertical Jump- standing vertical jump, in inches
- Broad Jump- the length, in inches of a player's standing broad jump
- Shuttle- time in a 20-yard back and forth "shuttle" run, in seconds
- 3cone- time in seconds taken to complete the 3-cone drill, another short-shuttle type drill involving running to various cones

3 Exploratory Data Analysis

3.1 Classification

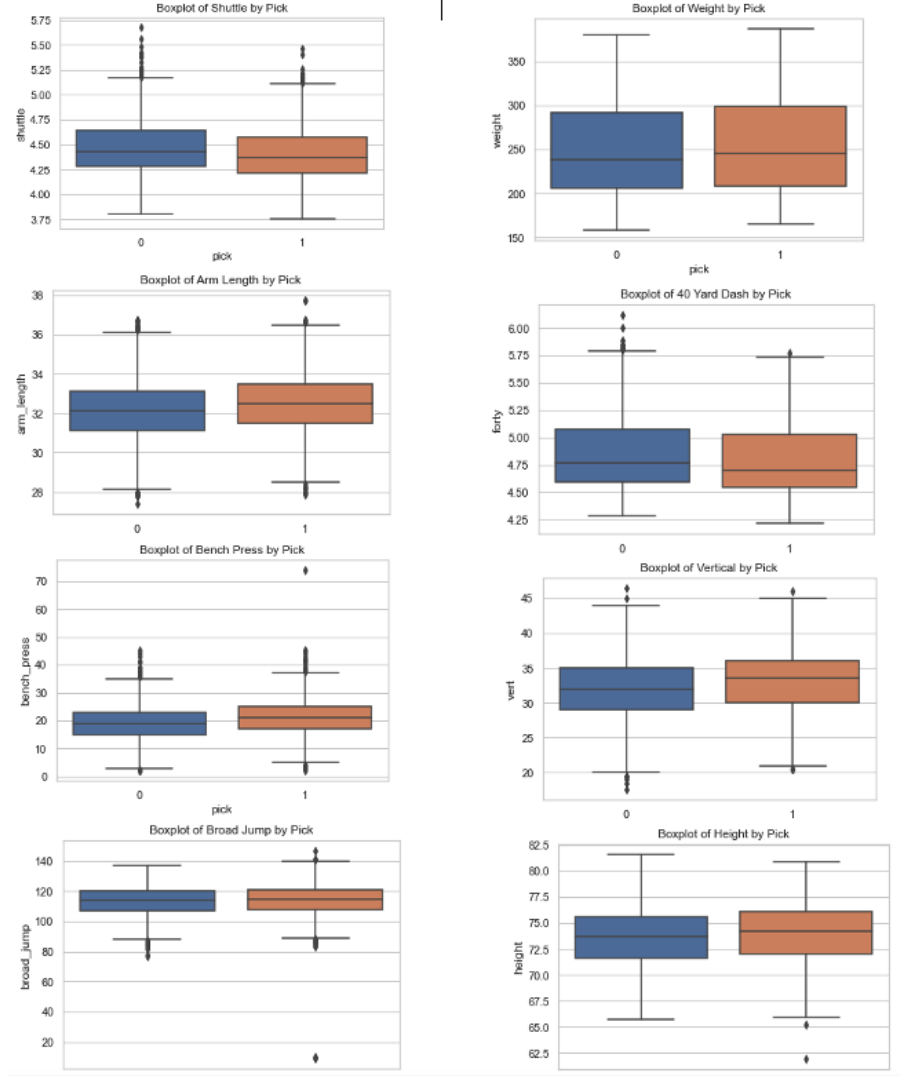


Figure 1: Predictor Boxplots: Drafted vs. Un-drafted

We convert the draft pick column to a binary variable where 0 indicates a player was un-drafted and 1 indicates the player was drafted after the combine. To get a general understanding of our data and a classification goal we observe drafted vs un-drafted statistics. From the pick column we see that about 58.6% of people who participated in the combine and 41.4% did not.

The boxplots in Figure 1 visualize the distributions of some of our numeric variables with respect to their draft standing (0 = un-drafted and 1 = drafted). From the boxplots we can see that variables such as Height and Broad Jump do not look very indicative of draft pick as the plots are very similar to one another. Other features such as 40-Yard Dash and Shuttle appear to have larger statistical differences between the two groups. From the 40-Yard Dash it is evident that players who are drafted tend to be quicker than those who are not, which is also supported by

	height	weight	hand_size	arm_length	forty	bench_press	vert	broad_jump	shuttle	cone
height	1.000000	0.749966	0.494092	0.723978	0.626795	0.357892	-0.438507	-0.424118	0.528578	0.508327
weight	0.749966	1.000000	0.496161	0.602941	0.883426	0.626132	-0.673852	-0.717905	0.748888	0.771973
hand_size	0.494092	0.496161	1.000000	0.504237	0.411424	0.294874	-0.275543	-0.274962	0.340291	0.346123
arm_length	0.723978	0.602941	0.504237	1.000000	0.468437	0.259697	-0.316546	-0.279479	0.461745	0.438580
forty	0.626795	0.883426	0.411424	0.468437	1.000000	0.459114	-0.754774	-0.795473	0.772295	0.809775
bench_press	0.357892	0.626132	0.294874	0.259697	0.459114	1.000000	-0.314744	-0.368405	0.374785	0.403321
vert	-0.438507	-0.673852	-0.275543	-0.316546	-0.754774	-0.314744	1.000000	0.786275	-0.691112	-0.668508
broad_jump	-0.424118	-0.717905	-0.274962	-0.279479	-0.795473	-0.368405	0.786275	1.000000	-0.670406	-0.702932
shuttle	0.528578	0.748888	0.340291	0.461745	0.772295	0.374785	-0.691112	-0.670406	1.000000	0.801462
cone	0.508327	0.771973	0.346123	0.438580	0.809775	0.403321	-0.668508	-0.702932	0.801462	1.000000

Figure 2: Numerical Predictor Correlation Matrix

the Shuttle plot. Additionally, there are significantly fewer "negative" outliers for drafted players in both features. The boxplots support the notion that players who are drafted tend to perform better in athleticism based tests in the NFL Combine, although the actual predictive power of these measurements is questionable.

We produce a feature correlation matrix shown in Figure 2 which unsurprisingly shows evidence of collinearity amongst various predictor pairs. Here shuttle, cone, forty, and weight are all highly correlated with each other while broad jump and vertical jump also show strong correlation. The first set of predictors are speed-related metrics, with the latter set both relating to jumping so these results are logical. Collinearity is an issue that both causes unreliable statistical inference and unnecessarily complicates models, and we therefore plan to address this issue when building out our models.

3.2 Regression

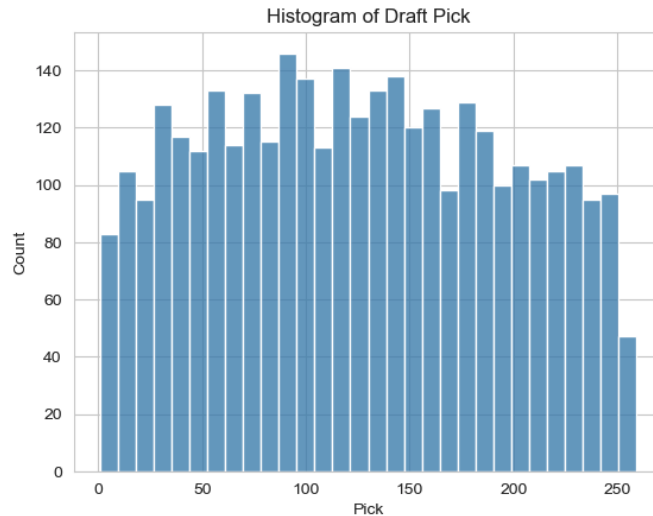


Figure 3: Draft Pick Distribution

The first thing we examine for the regression model is the distribution of the response variable displayed as a histogram in Figure 3. Although the distribution of actual draft picks is uniform, the histogram looks somewhat more normally distributed. This makes sense given the fact that many top draft picks do not want to hurt their prospects at the combine or at least not partake in certain drills, and players picked towards the end of the draft are sometimes not invited to the combine.

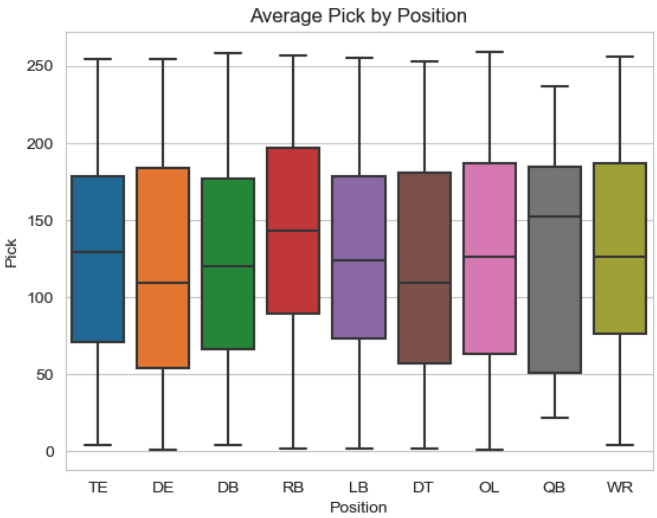


Figure 4: Average Pick by Position

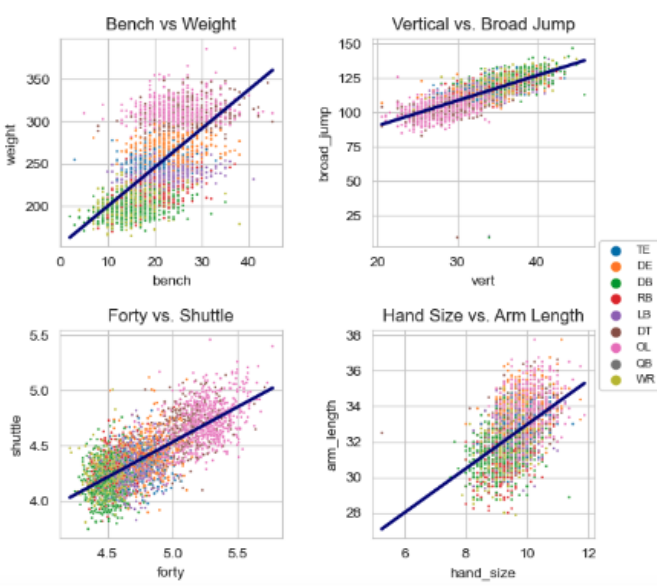


Figure 5: Comparing Predictor Variables

Figure 4 displays a boxplot of the categorical variable, position, with the response variable, pick. Overall, the positions are fairly similar to each other, with most positions having similar pick distributions. Quarterbacks have the most skewed distribution since a top QB is the most valuable position on the field while a middle-of-the-road quarterback holds very little value. Defensive Ends

and Tackles, who are often tasked with bringing down the quarterback on the opposing team, have the lowest average draft pick selection.

Similar to the full dataset, Figure 5 shows strong correlation amongst many of the predictors and therefore multi-collinearity needs to be addressed. Figure 6 shows the response variable plotted against four predictor variables that showed promise in the classification EDA. Surprisingly, none of the four variables appear to have a very strong relationship with pick. While these plots are somewhat discouraging, we still believe we can produce an interesting study even if our prior thoughts about combine results are disproven.

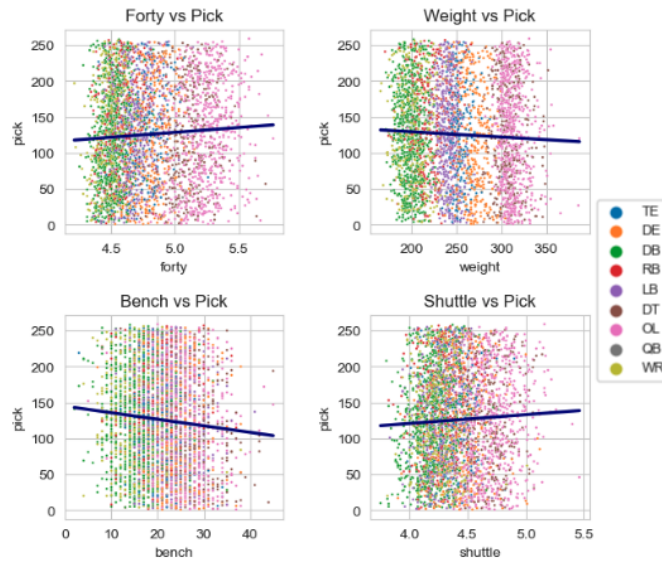


Figure 6: Comparing Predictors with Response



Figure 7: 40 time vs. Weight

Figure 7 displays the interaction between forty yard dash time and weight. The logic behind

including this plot is that players who are both big and fast would seemingly be picked towards the top of the draft. We encode the draft pick variable into categorical "draft rounds" for visual effect. Generally, the residuals confirm our intuition, particularly with first round picks. On average these top draft picks run faster (lower time) than other players of the same weight, and are bigger than players who run as fast.

4 Methodology and Results

4.1 Classification

4.1.1 Logistic Regression

We begin our classification with a simple Logistic Regression model containing every variable. In this initial model we use the variables Height, Weight, Hand Size, Arm Length, Forty, Bench Press, Vertical, Broad Jump, Shuttle, Cone, and Position against our response variable of whether or not a player was drafted. This simple and efficient Logistic model yields a cross validated (CV = 10) accuracy of 64.3%. The initial model correctly predicts 1,127 of our 1,753 observations. The model identifies 81% of drafted players at the expense of only correctly classifying 43% of un-drafted players.

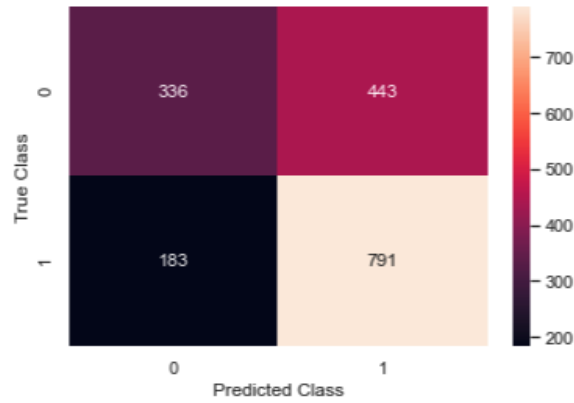


Figure 8: Full Model Logistic Regression Confusion Matrix

Using a feature importance plot from a subsequent Random Forest model we create a reduced model, eliminating features deemed statistically insignificant. The reduced model includes only forty, weight, and shuttle and attains a cross validated accuracy of 65.8%. The model increases accuracy by 1.5% while also significantly reducing model complexity. Seemingly then, combine results are a pretty good indicator of whether or not a player is drafted.

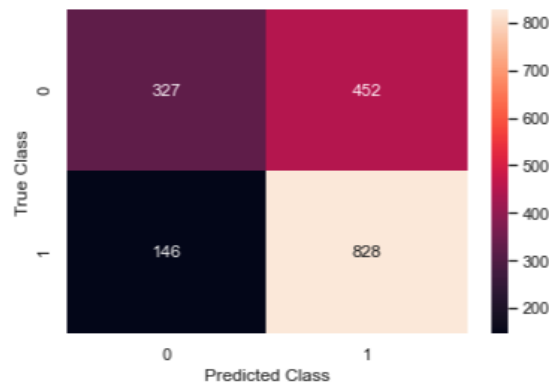


Figure 9: Reduced Model Logistic Regression Confusion Matrix

Figure 9 shows the cross-validated confusion matrix for the reduced logistic regression model while Figure 10 displays various other model metrics. From our reduced model we get coefficients of 0.1765 for hand size, 0.1667 for arm length, 0.0508 in bench press, and 0.1268 in vertical. This means that a 1 inch increase in hand size is associated with an increase of 19.3% in the odds of being drafted, a 1 inch increase in arm length is associated with an increase of 18.1% in the odds of being drafted, a 1 rep increase in bench press is associated with an increase of 5.2% in the odds of being drafted, and a 1 inch increase in vertical is associated with an increased of 13.5% in the odds of being drafted.

	precision	recall	f1-score	support
0	0.69	0.42	0.52	779
1	0.65	0.85	0.73	974
accuracy			0.66	1753
macro avg	0.67	0.63	0.63	1753
weighted avg	0.67	0.66	0.64	1753

Figure 10: Reduced Model Logistic Regression Metrics

In addition to being more accurate than the full model, the reduced model increased precision from 64% to 67% and recall from 64% to 66%. Although the reduced model outperforms the full model in nearly every statistic, the reduced model actually produces a worse recall score amongst un-drafted players at just 42%.

4.1.2 Random Forest

We follow a similar protocol as the logistic regression model by starting with all variables to fit a random forest model that produces an accuracy of 64.2%, which falls short of our logistic regression results.

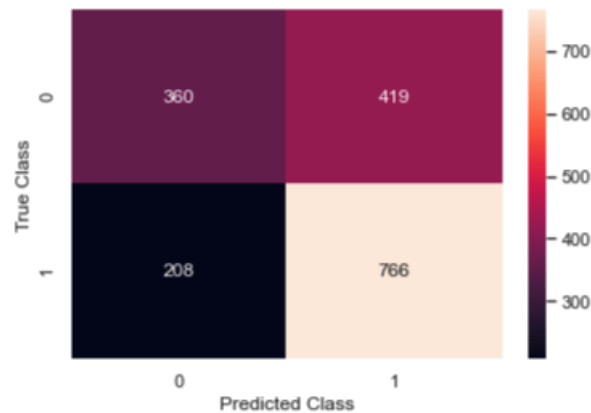


Figure 11: Random Forest Confusion Matrix

Figure 11 displays the cross-validated confusion matrix for the random forest model, while Figure 12 displays model metrics. Most of the results from the random forest fall slightly below our logistic models results. The one thing that stands out from the random forest model is the

	precision	recall	f1-score	support
0	0.63	0.46	0.53	779
1	0.65	0.79	0.71	974
accuracy			0.64	1753
macro avg	0.64	0.62	0.62	1753
weighted avg	0.64	0.64	0.63	1753

Figure 12: Random Forest Metrics

4% increase in recall amongst un-drafted players, meaning that it identifies about 4% more of the true un-drafted players. The trade-off here is precision, where the reduced logistic model is 6% more precise amongst un-drafted players.

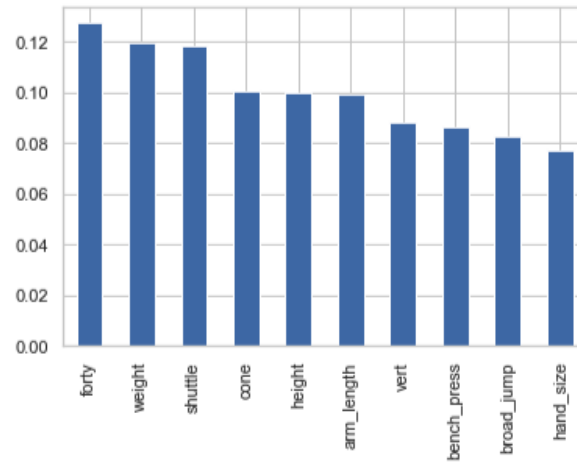


Figure 13: Random Forest Feature Importance

Figure 13 represents feature importance for the Random Forest Model with forty, weight, and shuttle being the three most important variables. We further attempt to optimize our model through variable reduction, and achieve a max accuracy of about 63.8% which falls just short of the full random forest model performance. Therefore we decide to keep the Random Forest Model as is and include each numerical predictor.

4.2 Regression

4.2.1 OLS

We begin our regression analysis with an OLS statistical model using draft pick as the response. In our first model we use all 10 numerical predictors along with the positional categorical variable. Full model results yield an R-squared of 0.088 and an adjusted R-squared of 0.084, which means that only 8.4% of the variance in draft pick is explained by this set of predictors. The lack of strong linear relationships in our EDA scatterplots makes this result unsurprising.

Next, we implement a backward selection process where we iteratively remove the max p-value variable until all remaining predictors are statistically significant. The final model contains 5 numerical variables and 3 of the 8 categorical dummy variables.

OLS Regression Results						
Dep. Variable:	pick		R-squared:	0.084		
Model:	OLS		Adj. R-squared:	0.082		
Method:	Least Squares		F-statistic:	39.27		
Date:	Fri, 03 Mar 2023		Prob (F-statistic):	2.81e-60		
Time:	07:25:40		Log-Likelihood:	-19210.		
No. Observations:	3418		AIC:	3.844e+04		
Df Residuals:	3409		BIC:	3.849e+04		
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-222.0558	49.765	-4.462	0.000	-319.628	-124.484
weight	-0.6549	0.088	-7.462	0.000	-0.827	-0.483
arm_length	-6.3900	1.051	-6.079	0.000	-8.451	-4.329
forty	119.8296	9.628	12.445	0.000	100.952	138.708
bench	-0.5226	0.249	-2.102	0.036	-1.010	-0.035
shuttle	37.9831	7.230	5.253	0.000	23.807	52.159
DB	-15.0889	3.667	-4.115	0.000	-22.279	-7.899
DT	-18.5718	5.954	-3.119	0.002	-30.245	-6.898
OL	-28.7441	5.753	-4.996	0.000	-40.024	-17.465
Omnibus:	486.318		Durbin-Watson:	1.925		
Prob(Omnibus):	0.000		Jarque-Bera (JB):	129.009		
Skew:	0.147		Prob(JB):	9.68e-29		
Kurtosis:	2.095		Cond. No.	1.13e+04		

Figure 14: OLS Summary

The model is summarized by Figure 14. In the final model, each individual variable is highly significant with the response variable, except for bench, which is only significant at the $\alpha = 0.05$ level. Not surprisingly, shuttle and forty, which relate to timed speed, are positively correlated with a later draft pick. This makes sense because lower forty and shuttle times indicate faster speed. Additionally, the other numerical predictors that relate to size are negatively correlated with draft pick which is also logical—bigger players, holding all else equal, are generally drafted earlier.

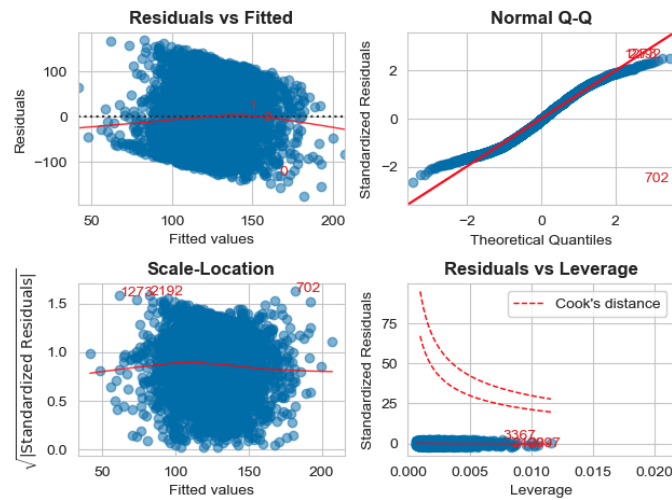


Figure 15: Diagnostic Plots

The diagnostic plots shown in Figure 15 bring into question the goodness-of-fit of a linear model. The errors do not appear to be normally distributed as shown by the residuals vs. fitted

and normal q-q plots. One positive note is that the data does not appear to contain any outliers, as shown by the scale-location and leverage plots. Unfortunately, re-specifying the model is not straightforward. Using a robust regression would not have any effect since the data does not contain outliers. Implementing polynomial regression or basis splines on the other hand, is complicated given the number of predictors utilized. The best approach is likely transforming the pick variable into a categorical variable with multiple rounds and then running a multinomial regression. However, since we already conduct a classification model in this study, we decide to leave the variable as is and proceed with a regression approach.

4.2.2 Predictive Linear Modeling

Next, we create a set of predictive models, cross-validate them and compare results in terms of RMSE. The first predictive model we build is based on the final OLS model. We conduct 3-fold cross validation and yield an RMSE of 67.02. The null model standard deviation is 69.8, meaning if you were to just predict the mean for each pick you would be off by 69.8 picks on average. Therefore the predictive power of the model is limited.

4.2.3 PCR

We implement PCA/PCR since our data consists of several features that are highly correlated with each other. Principal Component Analysis (PCA) is a method used to reduce predictor dimensions by eliminating collinearity within the dataset. The first principal component is computed by finding a linear combination of input variables with greatest variation in the data. Subsequent components are computed by finding the linear combination of variables with maximum variation in the data, that is also orthogonal to all prior components.

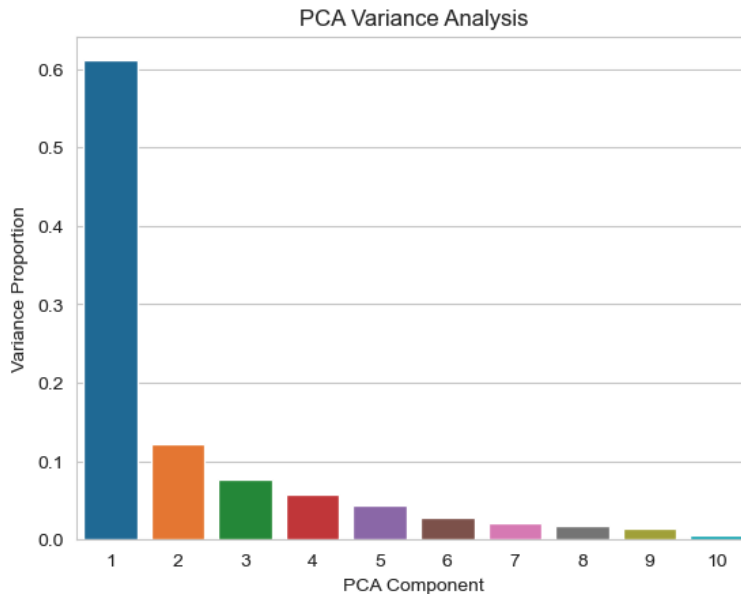


Figure 16: PCA Variance

Figure 16 shows PCA results for our 10 numerical predictors. The first principal component accounts for over 62% of the variation in the data and by the 5th principal component, 91.1% of variation in the data is accounted for.

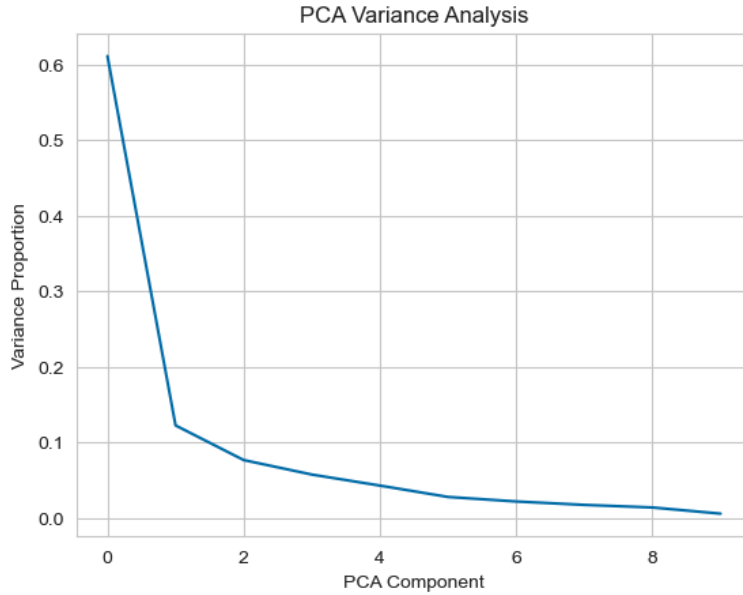


Figure 17: PCA Variance Analysis

Figure 17 displays total variance of each component. Prior to implementing PCR, we must determine how many principal components to use in the model. The rate change in variance proportion significantly flattens after both 2 and 5 principal components, and we therefore choose 3 Principle Components, which is between the two elbows.

	PC1	PC2	PC3
height	0.304736	-0.405880	-0.107401
weight	0.385545	-0.038562	0.142756
hand_size	0.228303	-0.461385	0.002039
arm_length	0.263038	-0.533021	-0.253794
forty	0.373551	0.149111	-0.022600
bench	0.232707	-0.048905	0.913638
vert	-0.319665	-0.333223	0.195778
broad_jump	-0.324417	-0.386599	0.068543
shuttle	0.342782	0.121846	-0.135350
3cone	0.344318	0.191882	-0.084137

Figure 18: PCA Loadings

Figure 18 shows PCA loadings for each of the three principal components we selected. Generally, we classify PC1 as a summary of a player's size. PC2 is more difficult to explain but appears inversely related to hand/arm size and jumping ability, while PC3 is almost purely strength related.

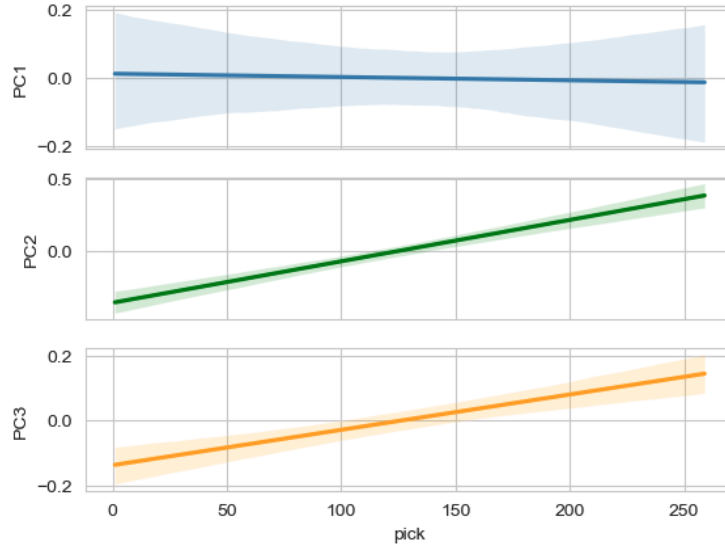


Figure 19: PCA vs Response

Figure 19 shows each principal component’s relationship with the response variable. The first principal component shows 0 correlation with the response variable y . This is not surprising since PCA does not take into account the response variable at all. The other two principal components show a much larger positive correlation with draft pick however, meaning the larger each component is the later in the draft a player will be selected. This makes a lot of sense in regards to PC2 given its loadings, but since PC3 is mostly composed of strength, a positive correlation is surprising.

To ensure that our PCR model is fairly measured against our other models, we add the one-hot encoded variables for the three positions used in the linear model: DB, DT and OL. Using 3-fold cross validation, the PCR model produces an RMSE of 68, meaning that on average, the model will predict a draft pick position that is 68 picks away from the actual value. Given a null model standard deviation of 69.8 this proves again to not be very predictive. The model also significantly underperforms compared to the linear model.

4.2.4 XGBoost

The final predictive model we use for the regression problem is XGBoost. Boosting is an ensemble method where committee trees are sequentially added to the prior committee. The idea behind boosting is to treat the error from the previous ensemble as the response variable for the next tree grown. XGBoost is a form of boosting where both L1 and L2 regularization are implemented into the loss function, and is solved using iterated re-weighted least squares.

XGBoost includes various hyper-parameters that can be adjusted. To implement hyper-parameter tuning, we first split the data into a training set containing 70% of observations and a testing set containing 30% of observations. We implement a cross-validated grid search on the training set to determine the optimal combination of `max_depth`, `eta`, and `n_estimators`. `Max depth` refers to the maximum depth, or layers in each tree that is built, `Eta` is the learning rate, and `n_estimators` is the number of trees that are built.

Cross-validation yields the following optimized parameters: `max_depth=1`, `eta=0,25`, and `n_estimators=100`. Finally, we run another iteration of XGBoost using the tuned hyper-parameters on the unseen test set. The result yields an RMSE of 68.45. Surprisingly the XGBoost model which is known to be quite powerful, performs worse than PCR and substantially worse than the linear model.

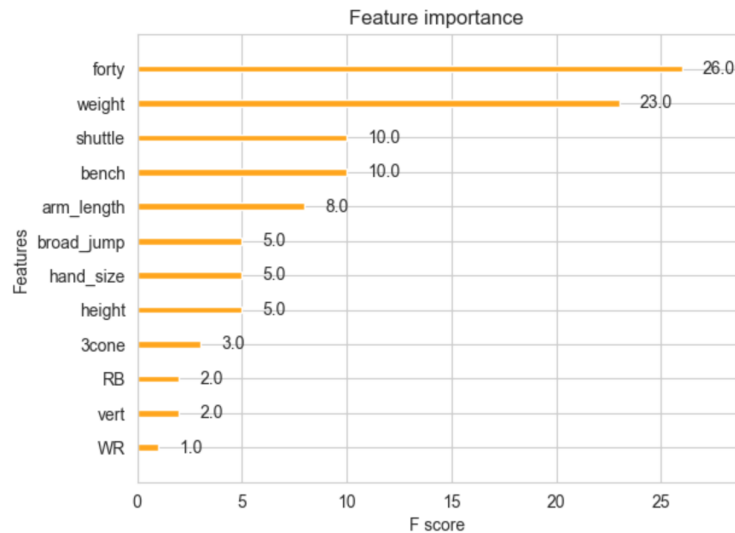


Figure 20: XGB Feature Importance

Figure 20 shows a plot of feature importance based on the F score of each variable in our XGBoost model. Forty time and weight are the two most predictive features in the dataset, with shuttle and bench belonging in the next tier. Speed, size and strength are generally important factors in playing football so these results seem logical. Not surprisingly these features also correspond with the variable importance in the OLS model.

5 Limitations, Further Research and Conclusion

5.1 Limitations

One limitation of the model is that we do not account for 1400 players who were drafted. The proportion of these players who did not attend the combine is unknown meaning we cannot rule out some mis-classifications. Additionally, we have significant amounts of missing data which limits the size of the dataset used. We also do not have any data on any interviews players partake in and cannot determine whether those interviews have any impact on draft position.

5.2 Further Research

One interesting thing that could be looked into is how players' combine stats compare to their performance once they actually reach the NFL. The question to be asked here would be: how predictive are combine stats in determining the quality of an NFL playing career? The previously mentioned missing data would be another important aspect to examine. Many players will only do certain drills at their school-specific Pro-Days, so using this data to impute missing combine data would be helpful in utilizing a larger dataset.

Another thing that could be examined is the small amount of data on the Wonderlic score, which is an "intelligence" test typically seen as being important for only certain positions. Finally, the data could be used to determine if a player is a candidate to play a different position in the NFL due to their combine numbers. For example, a player not projected to be drafted as a tight end due to lack of speed could have better odds to be drafted as an offensive lineman if they have enough size. It could be interesting for a player to explore potential positional transitions based on their physical metrics.

5.3 Conclusion

For the classification problem, the model with the the highest accuracy is the reduced-model Logistic Regression which produces an accuracy of 65.8% a metric we are satisfied with given the dataset. For the regression problem, the model that produces the lowest RMSE is the basic linear model using the subset of variables determined by the OLS model. Although the RMSE of 67.02 is not very predictive, the model does perform much better than both the PCR and XGBoost models.

Given the significantly better performance of the classification model compared to the regression model, we conclude that combine stats are more important in determining whether a player is drafted than where in the draft they are taken. Combine results are seemingly used more as a threshold a prospect must meet before other factors are used to determine where a player is actually selected. We also conclude that other factors are likely much more important in determining draft position such as college stats and level of playing competition.

Presentation Questions

Nicole: In the Regression model, what was your backward selection process?

Our backward selection process began with a full model using all 10 numerical variables and 8 dummy variables. We iteratively dropped the variable with the highest p-value until each variable had a p-value significant at the $\alpha = 0.05$ level, leaving us with our final model.

Yulin: Can you explain why your data has missing data and what your approach to this data was?

The data contained a lot of missing data because players will often only choose to partake in only certain drills at the combine. In many situations, players will only participate in running drills at their "Pro Day" which is basically a mini-combine done for each individual college. For our purposes, after removing two columns that contained more than 2/3rds NAs every variable outside of height and weight contained at least 10% missing values. At this point we decided to keep only the complete cases which represented about half of the data. Although throwing out half the data is not ideal, we still had about 6,000 observations to work with which is why we felt comfortable just using these observations.

References

- [1] Christina Gough. Total revenue of the national football league 2001-2021. *Statista*, January 2023.
- [2] NFL.com. Nfl scouting combine.
- [3] Nflcombineresults.com. Combine results, 2022.
- [4] Pro Football Reference. Historical draft picks, 2022.
- [5] Spotrac. Nfl draft salary per round, 2023.