Predicting Draft Status of College Linebackers

STSCI 4100 Project

Professor Yang

Nicholas Paschall (nbp33)

1. INTRODUCTION

Each year, the eagerly anticipated NFL Draft takes place, providing opportunities for college football players to live out their dreams, make millions of dollars, and help their new team win the Super Bowl. Many players seek to claim the lofty rewards brought by the Draft, but only few ever succeed. Every draft, there are approximately 16,000 draft-eligible players hailing from football programs across all levels of competition, yet only 1.6% of these players will continue their journey. Making it into the 1.6% largely depends on a player's production in college and their performance at the NFL Combine, an event showcasing their athleticism and skill set.

My intentions of this report are to build and analyze a model consisting of the most significant aspects of a player's game that contribute most to their chances of being drafted with the intention of making future predictions. Due to the variety of measurements and skillsets among the multitude of positions in football, it would be difficult to build a model that could accurately predict the probability of any one player to be drafted. So, for simplicity, I chose to strictly focus on the linebacker position.

2. DATA ACQUISITION PROCESS

2.1 Web Scraping

My data was scraped off the Pro Football Reference website, which maintains one of the largest databases for anything NFL related. This site offers a variety of information on every NFL Combine participant, including their college career, combine statistics, and whether they were drafted or not. I gathered the following variables of interest for each prospective player who participated in the NFL Combine along with their descriptive statistics as shown in **Table 1**. In total, 317 observations were collected consisting of 117 undrafted linebackers and 210 drafted linebackers. The distributions for each variable can be observed in **Appendix A**.

Table 1: Descriptive Statistics

Combine statistics highlighted in blue. Career statistics highlighted in orange. General statistics highlighted in green. P-values obtained using the Wald test against null intercept model.

Response	**Drafted (1), N (%)**		**Undrafted (0), N(%)**		**Total**
draft_fac: binary variable indicating if a player was drafted	210, 64.22%		117, 35.78%		327
Predictor	**Min**	**Median**	**Max**	**St. Dev.**	**p-value**
forty_yd: time in seconds a player sprints 40 yards	4.38	4.69	5.09	0.131	2.45E-10 ***
vertical: height in inches a player jumps off the ground	27	33.5	42.5	3.056	1.25E-06 ***
bench: number of reps a player successfully bench-presses 225 pounds	11	21	35	3.984	0.000231 ***
broad_jump: distance in inches a player jumps forward	104	119	139	5.854	8.54E-08 ***
cone: time in seconds a player runs the "3 cone-drill"	6.64	7.12	7.245	0.195	1.87E-06 ***
shuttle: time in seconds a player runs the "5-10-5" drill	4	4.31	4.96	0.135	1.97E-07 ***
tackles_solo: number of tackles player made unassisted	6	121	338	52.16	0.0494 *
tackles_ast: number of tackles player made assisted by fellow player	2	92	310	45.774	0.8079

tackles_tot: total number of solo and assisted tackles	8	218	633	91.386	0.229
tackles_loss: number of tackles player made causing a loss of yardage	0	22.5	74.5	11.37	0.00312 **
sacks: number of tackles player made on a quarterback before throwing a pass	0	7	37	6.612	0.00145 **
def_int: number of catches player made thrown by the opposing quarterback	0	1	14	2.117	0.465
pass_defended: number of incomplete passes caused by player	0	5	26	4.402	0.1255
forced_fumble: number of times player causes opposing ballcarrier to lose possession of ball	0	2	16	2.222	0.3228
Ht: height of player	5'10	6'2	6'6	0.306	0.0558
Wt: weight of player in pounds	202	238	259	9.032	0.000718 ***

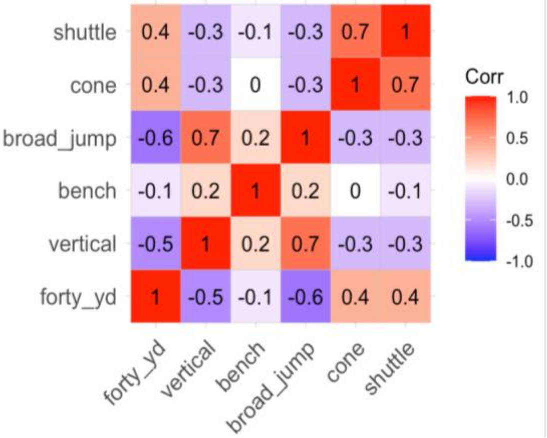
2.2 Data Cleaning

After completing the web scraping process, I noticed that many players opted out of specific combine events, and some players were lacking their college career statistics. This was particularly troubling when attempting to build a model due to the number of players being ignored with NA values. I implemented two strategies to resolve these issues.

First, within the Pro Football Reference site, each player who participated in the Combine would have a link connected to their college stats. For some players, this link was not available, which led to searching for their individual profiles on their respective teams. From there, I would gather the statistics of interest and insert them into the dataset.

Second, for players who had opted out of specific combine events, I chose to estimate their numbers based off other players with similar measurements. To begin the process, I first found the number of missing values for each event. The totals were as follows: 55 players opted out of the forty-yard dash, 89 players opted out of the vertical jump, 115 players opted out of the bench test, 81 players opted out of the broad jump, 177 players opted out of the 3-cone drill, and 169 players opted out of the shuttle drill. Next, the correlations between each measurement were computed as shown in **Figure 1**.

Figure 1: **Raw Combine Correlation Matrix**



With this information at my disposal, I was able to begin estimating. The idea of the estimation was to take one player with a missing value and obtain a small subset of similar players with non-missing values. I would average the non-missing values from this subset and apply it to the player with the missing value. The subset was obtained by using the variables most correlated with the variable of interest. As an example, say Player A opted out of the forty-yard dash, but has their broad jump and vertical measurements. Because broad jump and vertical are highly correlated with the forty-yard dash time, I would select players that had a vertical greater than or less than 0.5 inches compared to Player A's vertical, and players with a broad jump greater than or less than 3 inches compared to Player A's broad jump. I would use these remaining players' forty-yard dash times to estimate Player A's forty-yard dash time. Of course, this example works out because Player A had measurements for their vertical and broad jump while this was not the case for other players. Another issue was that there may have been no players that fit the specific boundaries for the vertical and broad jump, which resulted in having to expand these boundaries. Overall, there were numerous challenges faced to compute these estimations, which required many different trials and errors to produce the best results.

3. FEATURE SELECTION

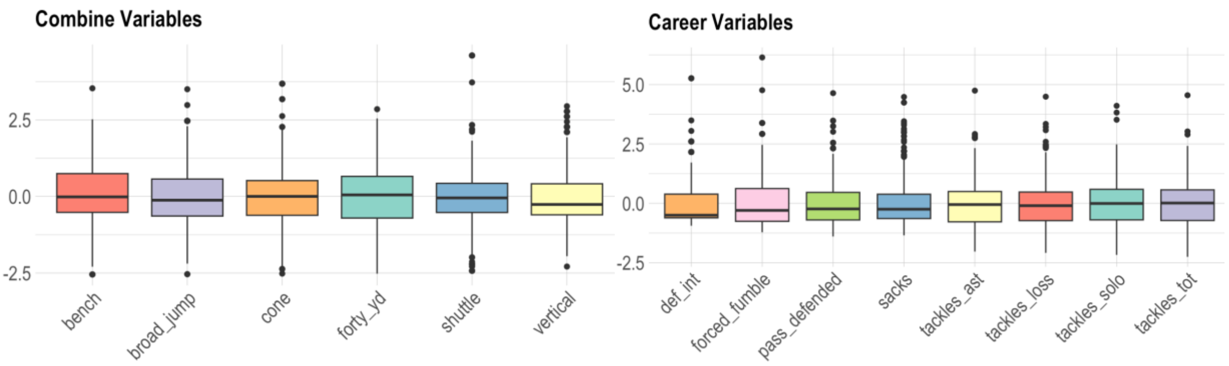
3.1 Data Preparation

Before diving into the feature selection, I employed upsampling to adjust the balance of our dataset by increasing the number of observations within the undrafted class. Upsampling involves randomly replicating observations from the minority class to match the number of observations in the majority class. This specific algorithm was chosen due to the dataset being on the smaller side with 327 total observations. With a 117:210 split between undrafted and drafted linebackers, this imbalance could lead to potentially biased models that perform classification poorly on the undrafted class. By increasing the representation of the undrafted class, the model can learn more effectively and make fairer predictions for both classes. After upsampling, the dataset grew to 420 total observations with 210 observations in the drafted and undrafted classes.

As a finishing touch, I also made sure to standardize each of the predictors to have a standard scale and zero mean. When predictors have different scales, variables with larger magnitudes can dominate the model's learning process and bias can result. Along with this, the coefficients associated also have different scales, making it challenging to compare their effects. By standardizing, this ensures that all variables are on a similar scale and allows for clearer understanding of each predictor's impact.

After standardizing and upsampling, the following box plots summarizing the distributions of each variable were produced in **Figure 2**.

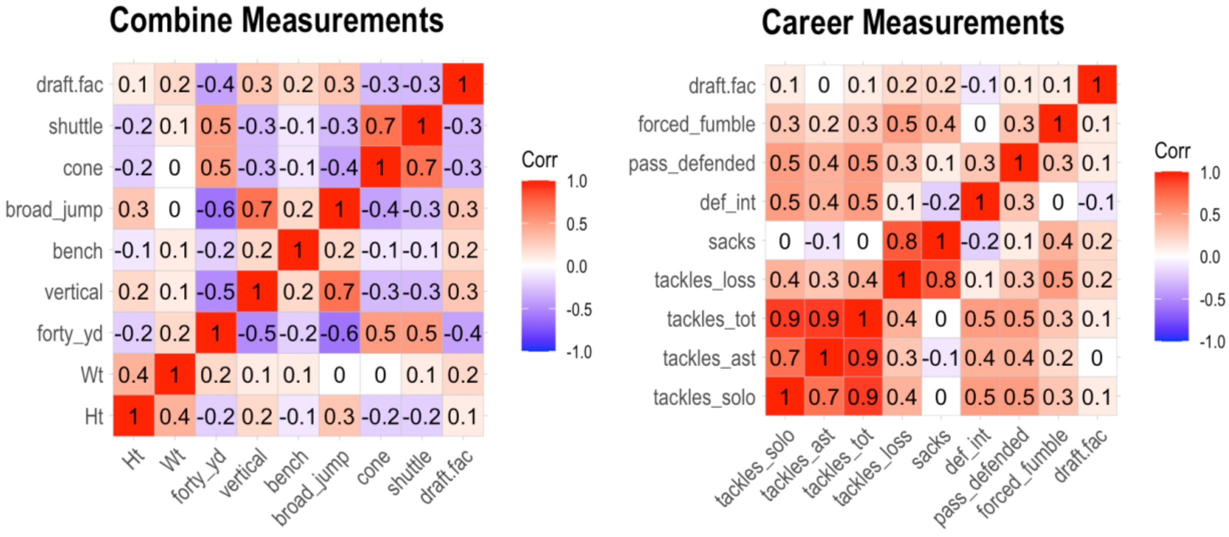
Figure 2: **Box Plots**



3.2 Correlation Plot

We begin the feature selection process by first looking at the correlation matrices for both the Combine and career variables in **Figure 3**. These matrices help gain a better understanding of which variables have significant effects on our response as well as identifying multicollinearity between variables.

Figure 3: **Correlation Matrices**



Regarding the Combine measurements, *height*, *weight*, *vertical*, *bench*, and *broad_jump* are positively correlated with the probability of a linebacker being drafted, whereas *shuttle*, *cone*, and *forty_yd* are negatively correlated. The most notable correlations with being drafted are *forty_yd*, *vertical*, *broad_jump*, *cone*, and *shuttle* which exhibit correlations of -0.4, 0.3, -0.3, 0.3, and -0.3 respectively.

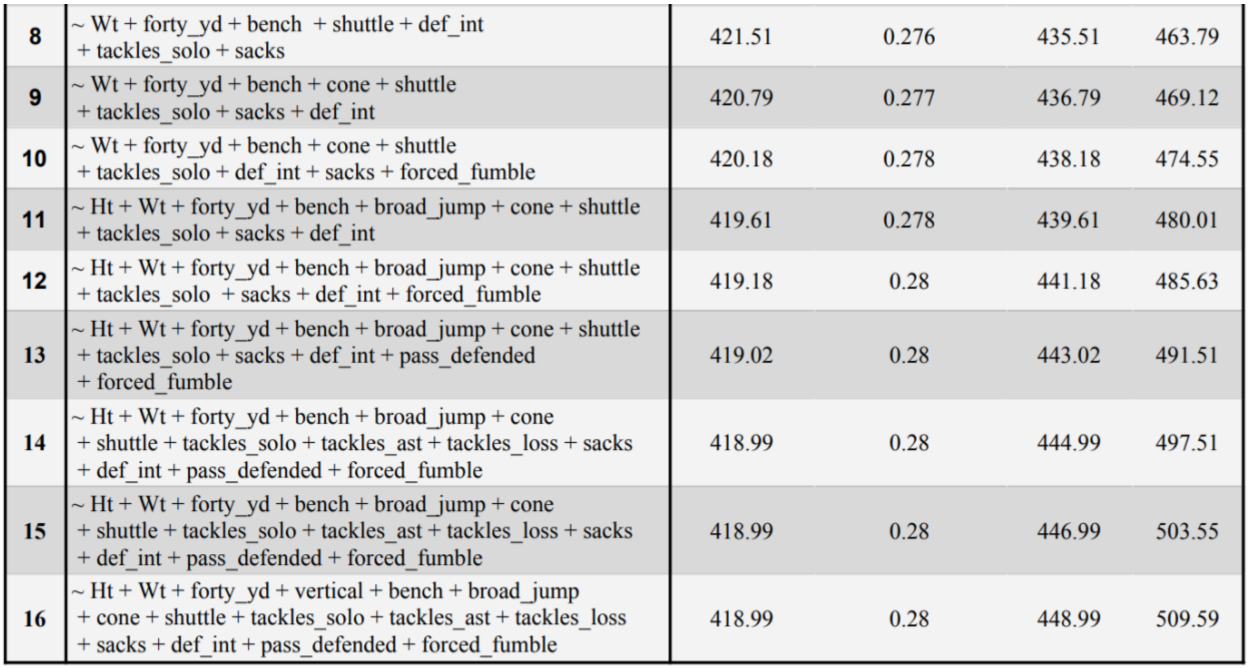
As for the career statistics, *tackles_solo*, *tackles_tot*, *tackles_loss*, *sacks*, *pass_defended* and *forced_fumble*, each present positive correlations with the probability of a linebacker being drafted, while *tackles_ast* shows no correlation, and *def_int* shows a negative correlation. Overall, the career statistics all show weak correlation with the response, but the most notable variables appear to be *tackles_loss* and *sacks*, which have correlations of 0.2. We can also see a case of multicollinearity between *tackles_tot*, *tackles_ast*, and *tackles_solo*. This makes sense due to *tackles_tot* being the sum of both *tackles_ast* and *tackles_solo*, so we remove the variable for the remainder of the analysis.

3.3 Best Subset Selection

To identify the significant variables and begin building our model for predicting the draft status of a linebacker, I chose the best subset selection algorithm. The best subset algorithm is ideal for variable selection and model building due its capabilities for picking the best models among all 2^p models. **Table 2** displays the best model for each *k* predictors along with their corresponding values for deviance, McFadden's R^2, AIC, and BIC.

Table 2: **Best Subset Selection**

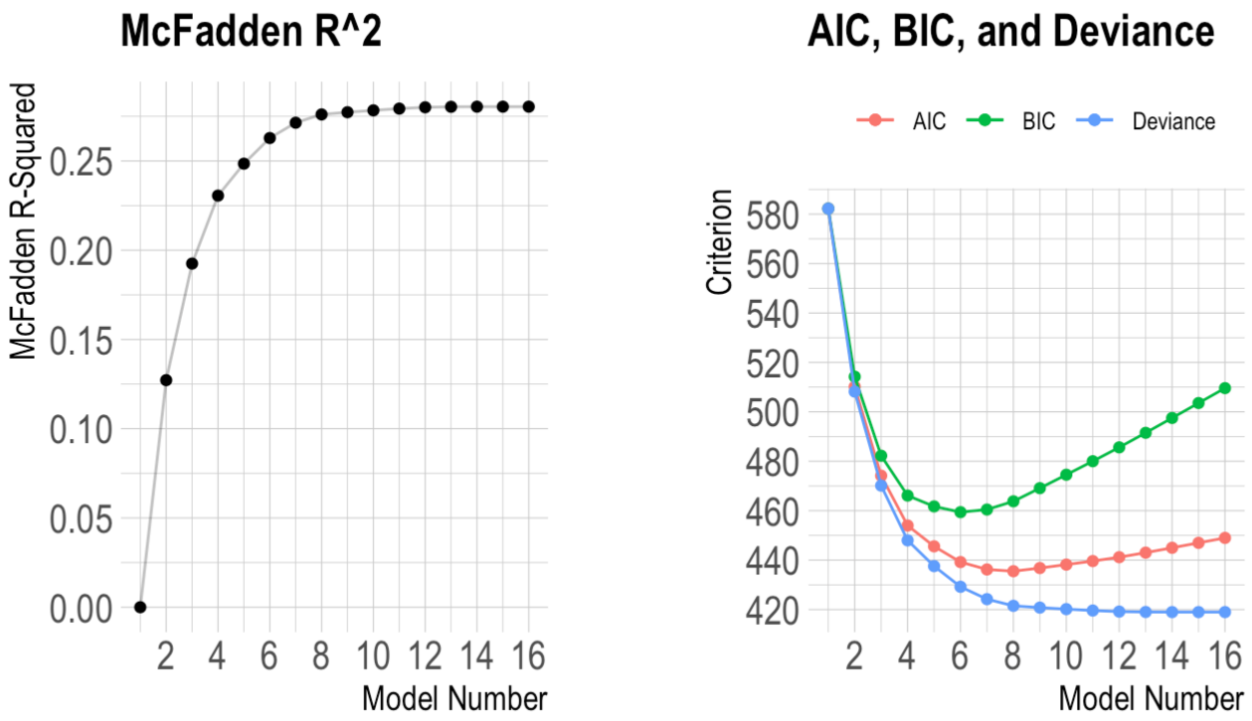
#	Model	Deviance	McFadden's R^2	AIC	BIC
1	~ 1	582.24	0	582.24	582.24
2	~ forty_yd	508.19	0.127	510.19	514.23
3	~ Wt + forty_yd	470.15	0.193	474.15	482.23
4	~ Wt + forty_yd + tackles_solo	448	0.231	454	466.11
5	~ Wt + forty_yd + shuttle + tackles_solo	437.57	0.248	445.57	461.73
6	~ Wt + forty_yd + bench + shuttle + tackles_solo	429.21	0.263	439.21	459.41
7	~ Wt + forty_yd + bench + shuttle + tackles_solo + sacks	424.22	0.271	436.22	460.46



8	~ Wt + forty_yd + bench + shuttle + def_int + tackles_solo + sacks	421.51	0.276	435.51	463.79
9	~ Wt + forty_yd + bench + cone + shuttle + tackles_solo + sacks + def_int	420.79	0.277	436.79	469.12
10	~ Wt + forty_yd + bench + cone + shuttle + tackles_solo + def_int + sacks + forced_fumble	420.18	0.278	438.18	474.55
11	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int	419.61	0.278	439.61	480.01
12	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int + forced_fumble	419.18	0.28	441.18	485.63
13	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int + pass_defended + forced_fumble	419.02	0.28	443.02	491.51
14	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	418.99	0.28	444.99	497.51
15	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	418.99	0.28	446.99	503.55
16	~ Ht + Wt + forty_yd + vertical + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	418.99	0.28	448.99	509.59

Figure 4 displays the AIC, BIC, Deviance, and McFadden R^2 values associated with each model. We can see that each of these values are either minimized or begin to level out at model 6 and beyond, but it is difficult to make a conclusion on the single best model based on these measurements.

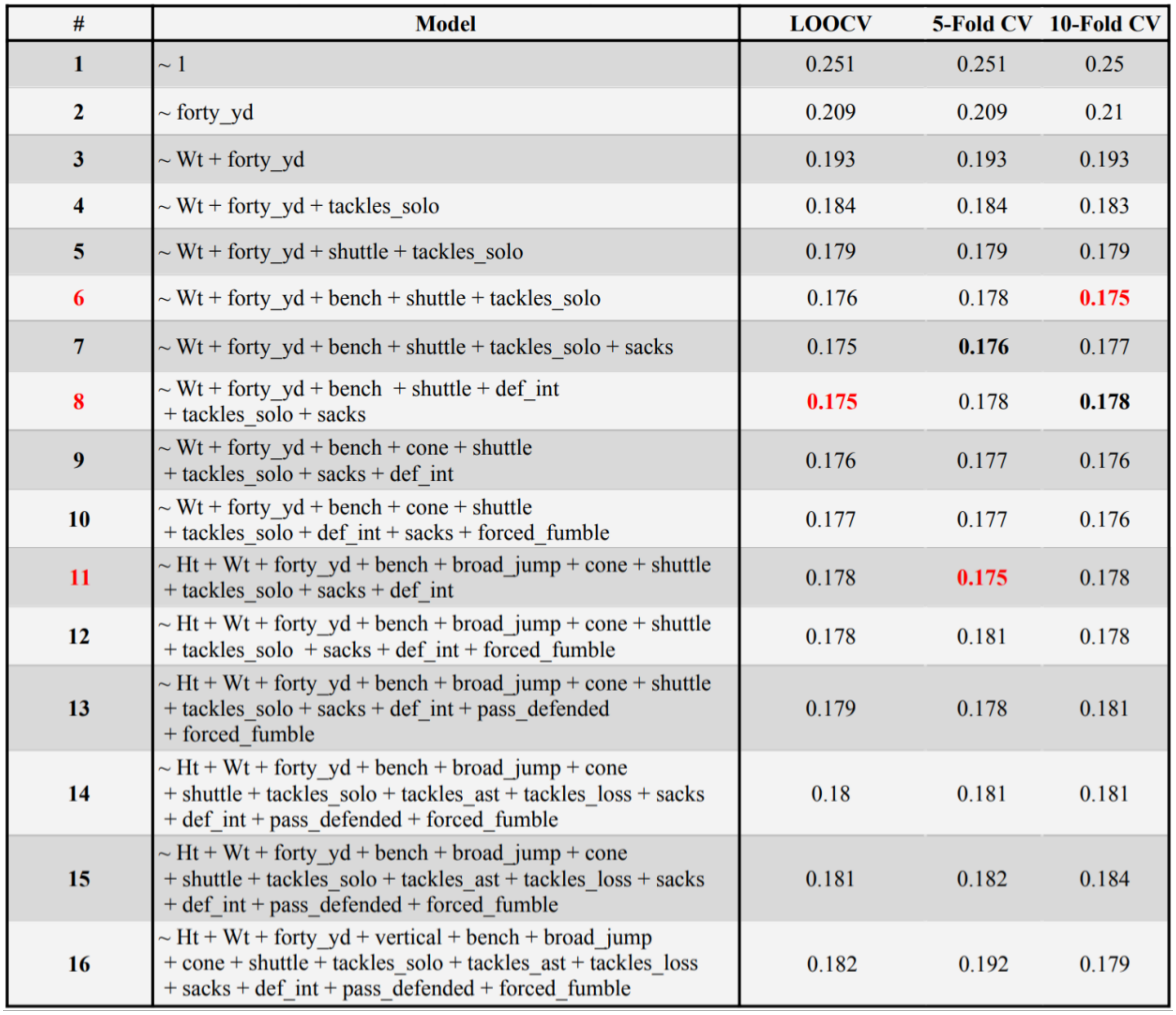
Figure 4: **Criterion Measurements**



3.4 Cross Validation

Because the main objective is to predict the probabilities of a linebacker being drafted, I used cross validation to help make the final decision regarding our model selection. Cross validation is a valuable tool for estimating a model's prediction performance while accounting for overfitting and offers several algorithms for balancing the bias-variance tradeoff to produce a trustworthy result. We utilized three algorithms: leave-one-out cross validation, 5-fold cross validation, and 10-fold cross validation. **Table 3** presents the validation errors produced from each of these methods for each model.

Table 3: **Cross Validation Results**



#	Model	LOOCV	5-Fold CV	10-Fold CV
1	~ 1	0.251	0.251	0.25
2	~ forty_yd	0.209	0.209	0.21
3	~ Wt + forty_yd	0.193	0.193	0.193
4	~ Wt + forty_yd + tackles_solo	0.184	0.184	0.183
5	~ Wt + forty_yd + shuttle + tackles_solo	0.179	0.179	0.179
6	~ Wt + forty_yd + bench + shuttle + tackles_solo	0.176	0.178	**0.175**
7	~ Wt + forty_yd + bench + shuttle + tackles_solo + sacks	0.175	**0.176**	0.177
8	~ Wt + forty_yd + bench + shuttle + def_int + tackles_solo + sacks	**0.175**	0.178	**0.178**
9	~ Wt + forty_yd + bench + cone + shuttle + tackles_solo + sacks + def_int	0.176	0.177	0.176
10	~ Wt + forty_yd + bench + cone + shuttle + tackles_solo + def_int + sacks + forced_fumble	0.177	0.177	0.176
11	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int	0.178	**0.175**	0.178
12	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int + forced_fumble	0.178	0.181	0.178
13	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + sacks + def_int + pass_defended + forced_fumble	0.179	0.178	0.181
14	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	0.18	0.181	0.181
15	~ Ht + Wt + forty_yd + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	0.181	0.182	0.184
16	~ Ht + Wt + forty_yd + vertical + bench + broad_jump + cone + shuttle + tackles_solo + tackles_ast + tackles_loss + sacks + def_int + pass_defended + forced_fumble	0.182	0.192	0.179

As seen in **Table 3**, the model with 8 predictors has the lowest leave-one-out cross validation error, the model with 11 predictors has the lowest 5-fold cross validation error, and the model with 6 predictors has the lowest 10-fold cross validation error. Based on these results, I decided our final model will be the model consisting of six predictors. Not only does it have the lowest validation error and simplest model, but 10-Fold cross validation offers the lowest bias of the three methods while maintaining low variance as well. This makes for better model evaluations and predictions on new data.

4.5 Prediction

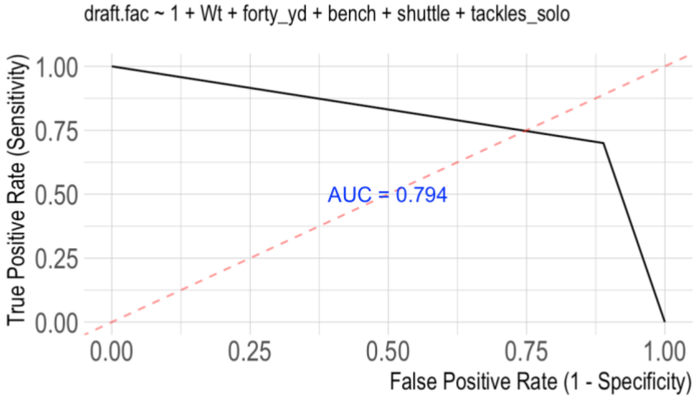
Lastly, I did a simple prediction assessment of our final model. The model is trained using the observations prior to 2022 to predict the draft status of those in 2022. **Table 4** presents the corresponding confusion matrix, which uses a threshold of greater than 0.5 to indicate a player being drafted. **Figure 5** shows the corresponding ROC plot for the model with an AUC of 0.794.

Table 4: **Confusion Matrix**

	Drafted	
Predicted	No	Yes
No	16	1
Yes	6	15

Test Error Rate: 18.4%

Figure 5: **ROC Curve**



4. INTERPRETATIONS

Using the whole, unstandardized, dataset to train our final model, we arrive at the following expression for predicting the probability of a linebacker being drafted:

$$draft.fac_i = 27.9033 + 0.091(Wt)_i - 7.7742(forty\_yd)_i + 0.0759(bench)_i - 3.492(shuttle)_i + 0.0095(tackles\_solo)_i + e_i$$

5.1 Coefficient Estimates and Interpretations:

The estimates and odds for the predictor coefficients are presented in **Table 5** along with the unit change used to calculate the odds.

Table 5: **Estimates and Odds**

Predictor	Estimate	Odds	Odds CI	Δ Change
Wt	0.091	1.095	(1.06, 1.14)	1
forty_yd	-7.7742	0.925	(0.90, 0.95)	0.01
bench	0.0759	1.079	(1.00, 1.16)	1
shuttle	-3.492	0.966	(0.94, 0.99)	0.01
tackles_solo	0.0095	1.01	(1.00, 1.02)	1

- A 1-unit increase in *Wt* is associated with a 6-14% increase in odds of being drafted.
- A 0.01-unit increase in *forty_yd* is associated with a 5-10% decrease in odds of being drafted.
- A 1-unit increase in *bench* is associated with a 0-16% increase in odds of being drafted.

- A 0.01-unit increase in *shuttle* is associated with a 1-6% decrease in odds of being drafted.
- A 1-unit increase in *tackles_solo* is associated with a 0-2% increase in odds of being drafted.

5. FINAL THOUGHTS

6.1 Issues

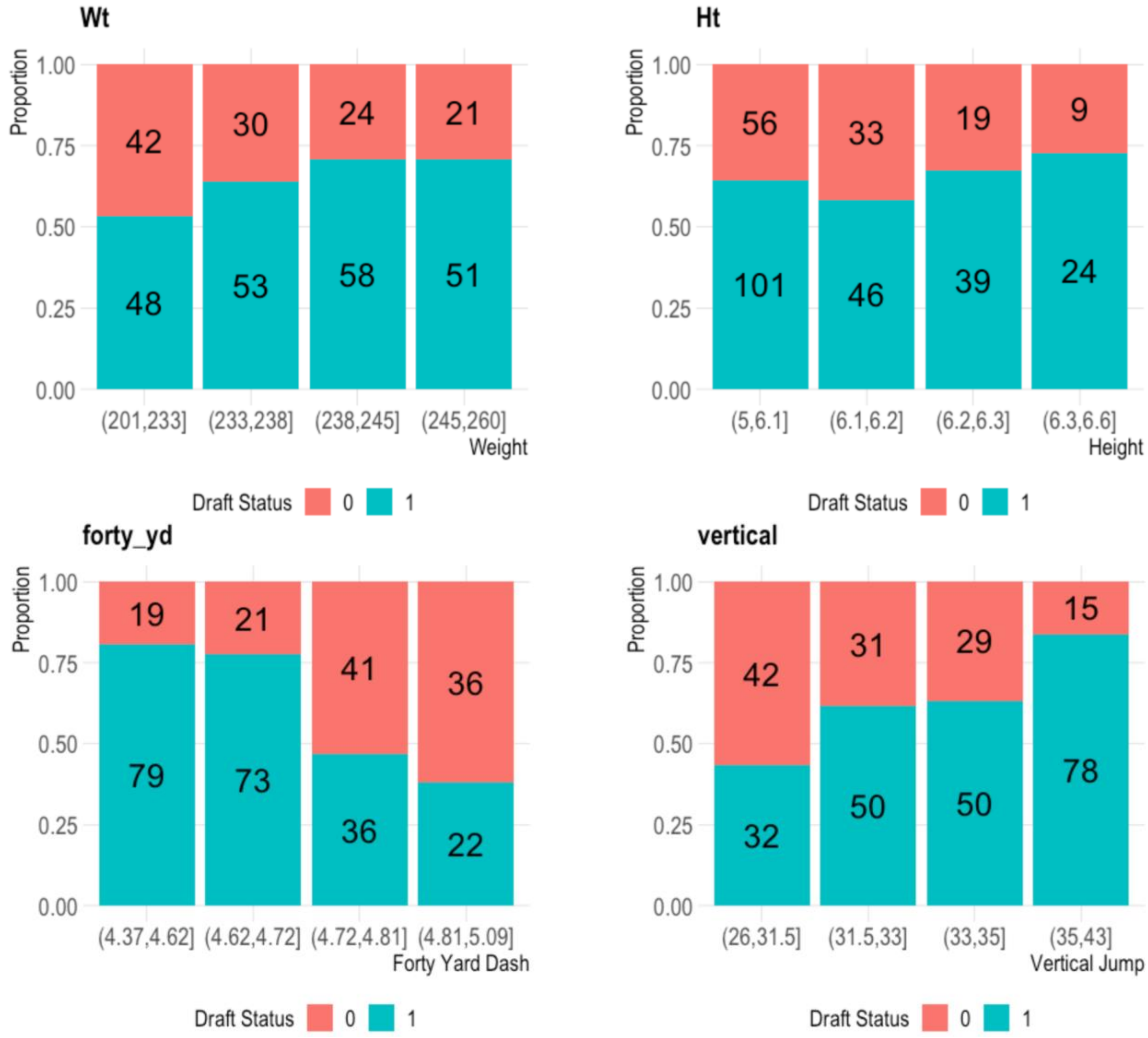
1. The abundance of estimations for NA values with regards to the Combine events is concerning and could possibly have some unknown impact on our results. A potential solution would be intensive exploration on players participating in these events elsewhere, such as, during a program's Pro Day.
2. Career stats can be a subjective matter to whoever is responsible for recording the stats for players each game. Some football programs, such as Cornell University, will not have a say in a player's stats if the opposing team is responsible for the official statistics of the game. Although we don't believe this problem is a major concern to our results, it is still something to consider.
3. More data!

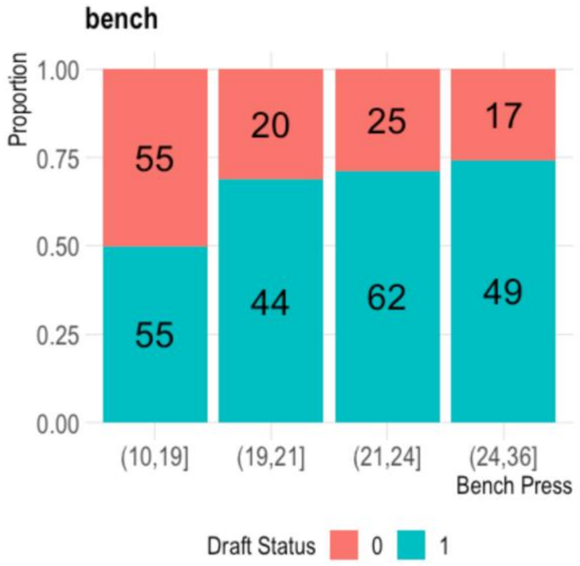
6.2 Potential Improvements

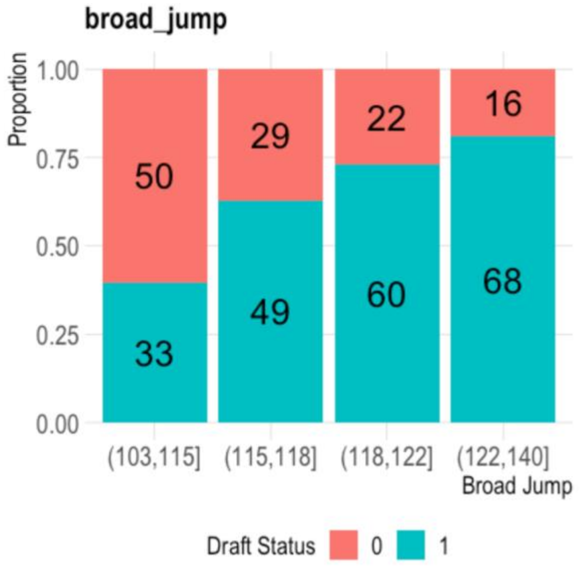
1. Accounting for a player's school, conference, and average strength of schedule
2. Accounting for a player's recent injuries relative to the Draft date and/or start of NFL season.
3. Accounting for NFL team needs and the demand for the position.
4. Weighting a player's career by year

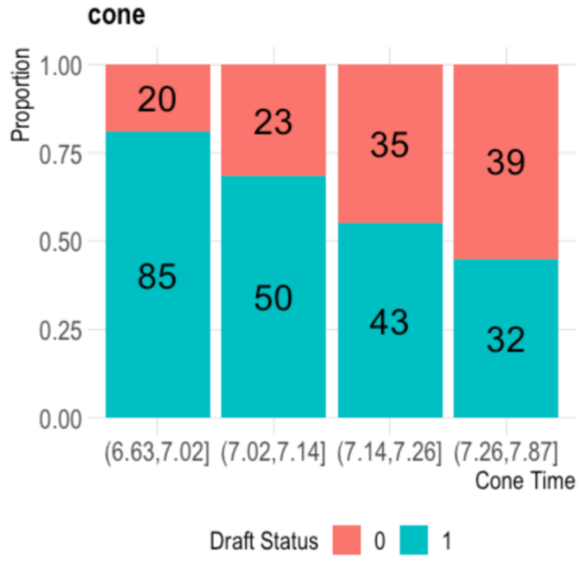
APPENDIX

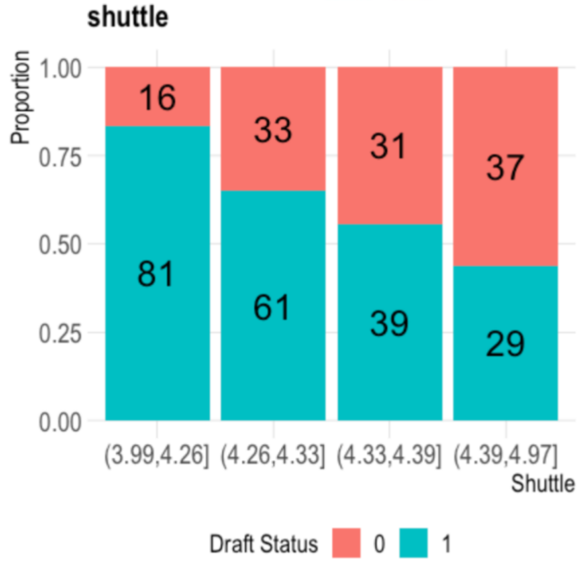
APPENDIX A: Box Plots for Each Variable

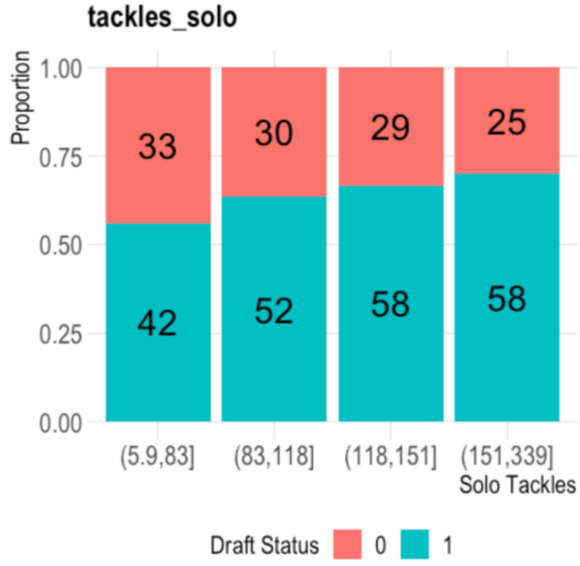


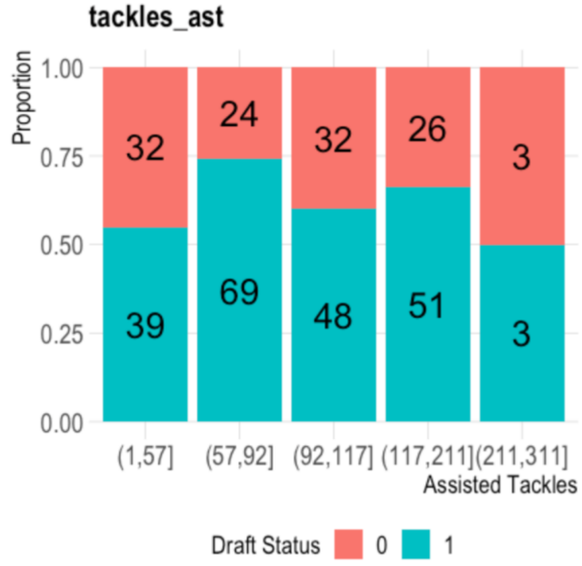

bench
Proportion
1.00
0.75
0.50
0.25
0.00
55
55
20
44
25
62
17
49
(10,19]
(19,21]
(21,24]
(24,36]
Bench Press
Draft Status
0
1

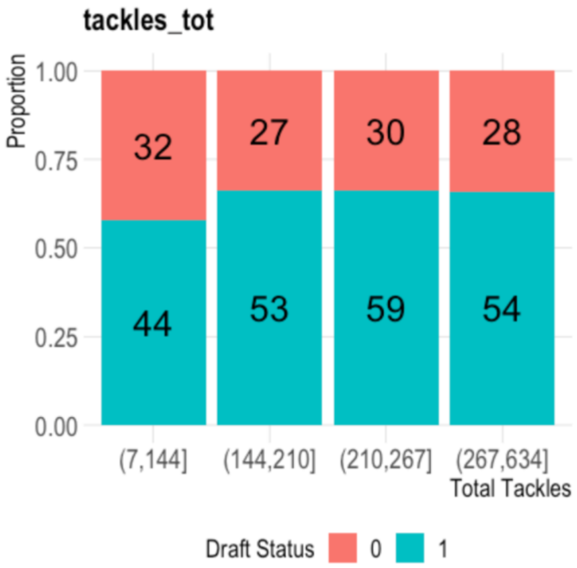

broad_jump
Proportion
1.00
0.75
0.50
0.25
0.00
50
33
29
49
22
60
16
68
(103,115]
(115,118]
(118,122]
(122,140]
Broad Jump
Draft Status
0
1

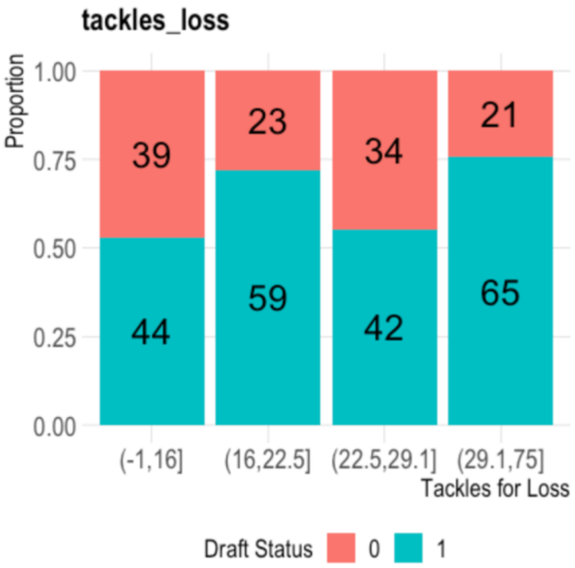

cone
Proportion
1.00
0.75
0.50
0.25
0.00
20
85
23
50
35
43
39
32
(6.63,7.02]
(7.02,7.14]
(7.14,7.26]
(7.26,7.87]
Cone Time
Draft Status
0
1

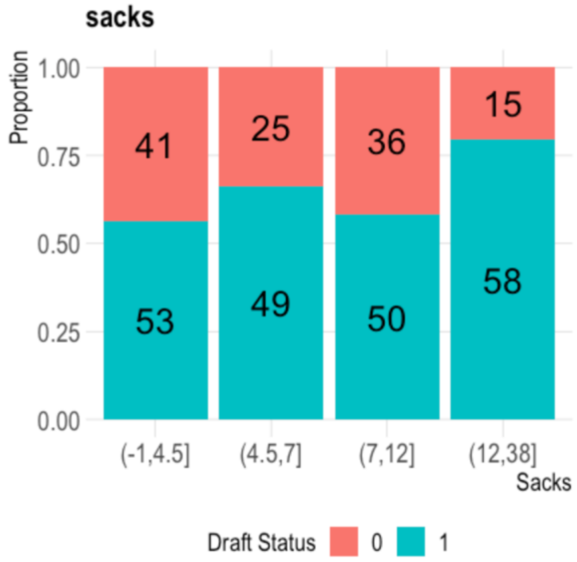

shuttle
Proportion
1.00
0.75
0.50
0.25
0.00
16
81
33
61
31
39
37
29
(3.99,4.26]
(4.26,4.33]
(4.33,4.39]
(4.39,4.97]
Shuttle
Draft Status
0
1

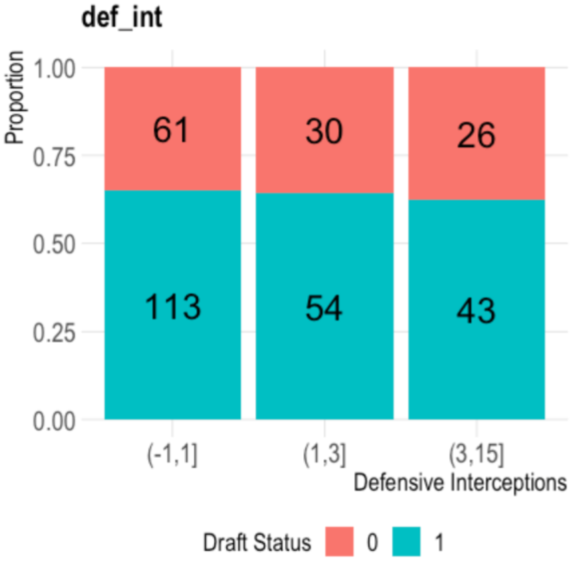

tackles_solo
Proportion
1.00
0.75
0.50
0.25
0.00
33
42
30
52
29
58
25
58
(5.9,83]
(83,118]
(118,151]
(151,339]
Solo Tackles
Draft Status
0
1

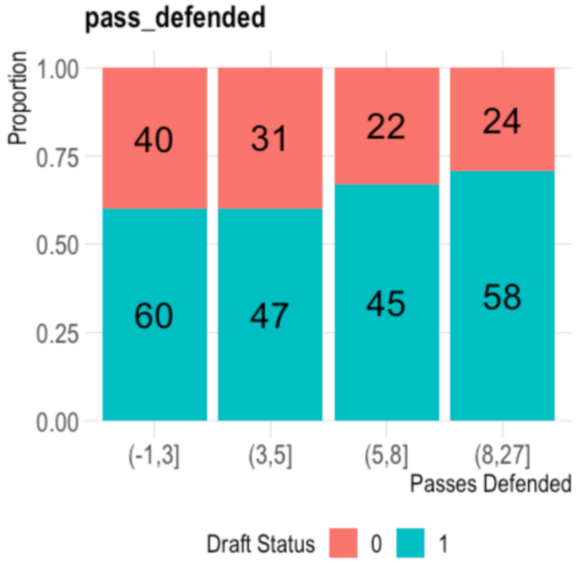

tackles_ast
Proportion
1.00
0.75
0.50
0.25
0.00
32
39
24
69
32
48
26
51
3
3
(1,57]
(57,92]
(92,117]
(117,211]
(211,311]
Assisted Tackles
Draft Status
0
1

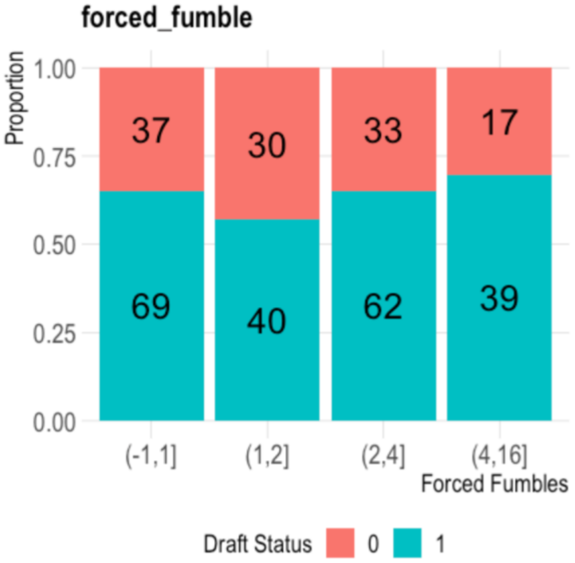

tackles_tot
Proportion
1.00
0.75
0.50
0.25
0.00
32
27
30
28
44
53
59
54
(7,144]
(144,210]
(210,267]
(267,634]
Total Tackles
Draft Status
0
1


tackles_loss
Proportion
1.00
0.75
0.50
0.25
0.00
39
23
34
21
44
59
42
65
(-1,16]
(16,22.5]
(22.5,29.1]
(29.1,75]
Tackles for Loss
Draft Status
0
1


sacks
Proportion
1.00
0.75
0.50
0.25
0.00
41
25
36
15
53
49
50
58
(-1,4.5]
(4.5,7]
(7,12]
(12,38]
Sacks
Draft Status
0
1


def_int
Proportion
1.00
0.75
0.50
0.25
0.00
61
30
26
113
54
43
(-1,1]
(1,3]
(3,15]
Defensive Interceptions
Draft Status
0
1


pass_defended
Proportion
1.00
0.75
0.50
0.25
0.00
40
31
22
24
60
47
45
58
(-1,3]
(3,5]
(5,8]
(8,27]
Passes Defended
Draft Status
0
1


forced_fumble
Proportion
1.00
0.75
0.50
0.25
0.00
37
30
33
17
69
40
62
39
(-1,1]
(1,2]
(2,4]
(4,16]
Forced Fumbles
Draft Status
0
1