Data Science & Analytics Final Paper

Data Collection

The data model that I built is a regression model that attempts to analyze how NFL Combine performance can affect a prospect's draft stock. To build the model, I collected data from Pro-Football-Reference. The raw data used to build the model was collected as two separate data sets. The first set contained draft data from 19 years of NFL Draft selections (2000-2018), including player name, position, year, round, and overall pick. The second set contained combine result data from all the players who participated in the events each year during the same time span. Once the data was imported into Excel, I merged the two data sets using the identifier of "Name" to complete the training data for the model. Below is a sample from the training data set:

Name	Position	Year	Round	Pick	40yd	Vertical	Bench	Broad Jump	3Cone	Shuttle
Baker Mayfield	QB	2018	1	1	4.84	29.0		111	7.00	4.28
Saquon Barkley	RB	2018	1	2	4.40	41.0	29			4.24
Sam Darnold	QB	2018	1	3	4.85	26.5		105	6.96	4.40
Denzel Ward	DB	2018	1	4	4.32	39.0	16	136		
Bradley Chubb	EDGE	2018	1	5	4.65	36.0	24	121	7.37	4.41
Quenton Nelson	C/OG	2018	1	6		26.5	35	105	7.65	4.62
Josh Allen	QB	2018	1	7	4.75	33.5		119	6.90	4.40
Roquan Smith	LB	2018	1	8	4.51					
Mike McGlinchey	OT	2018	1	9		28.5	24	105		
Josh Rosen	QB	2018	1	10	4.92	31.0		111	7.09	4.28

The Model / Analysis

I decided to pursue a regression model for this analysis. This modeling technique seeks to find relationships between independent variables (inputs) and a dependent variable (output). In doing so, the regression produces a formula that uses the historical data to predict what the future outputs might be. For this analysis, I ran a regression test with each of the six

combine events as independents and the overall selection in the draft as the dependent. One problem I ran into was that a regression cannot be run unless every column contains an equal number of variables. Since some players choose to participate in certain events at the combine but not all of them, it was not possible to include every player in the same regression test that featured all six events as independent variables. To solve this problem, I ran a unique regression test for each combine event, that way every player who participated in each specific event over the past 19 years was included in the analysis for a more accurate model. This produced formulas that could help predict where a player might be drafted based on his performance. This process was executed for all six events for all ten position groups (quarterback, running back, edge rusher, etc.) to generate the most accurate projections possible pertaining to the effects of combine results on draft stock, relative to each position group.

The R^2 Feature

An important output of a regression test is the R^2 value. This number signifies what percentage of the variance in the dependent variable can be explain by the independent variables. In other words, it represents the magnitude of the independent variable's effect on the dependent variable. In this regression analysis, the R^2 values describe how important each combine event is to a player's draft stock, relative to his position. For example, when running a regression to test the effect of the 40-yard dash on the draft stock of running backs, the R^2 was 0.118663 (otherwise known as 11.87%). This differs greatly from the R^2 of the regression testing the effect of the bench press on the draft stock of a running back, which is only 0.000707

(0.07%). From these outputs, it can be interpreted that the 40-yard dash is significantly more impactful on the draft stock of a running back than the bench press.

This aspect of the model can be very valuable to players preparing for the NFL Combine, because it can tell them which events will impact their draft stock the most. Keeping with the running back example, the R^2 values tell backs that they should focus on the 40-yard dash (11.87%), broad jump (10.56%), and vertical jump (6.22%) events as opposed to the 3-cone drill (3.96%), shuttle (0.20%), and bench press (0.07%). This chart displays which events are the most significant for each positional group, green being the most and red being the least:

Most Important Events by Position

	40yd	Vertical	Bench	Broad Jump	3Cone	Shuttle
C/OG	2.92%	1.38%	0.49%	3.59%	0.54%	0.31%
DB	11.41%	4.74%	0.12%	8.05%	4.75%	4.20%
DT	1.56%	0.21%	4.03%	0.86%	0.70%	0.71%
EDGE	5.01%	0.10%	0.73%	2.70%	2.50%	1.42%
LB	7.44%	6.45%	0.60%	8.46%	0.97%	1.62%
OT	9.11%	3.77%	5.31%	5.93%	2.27%	1.89%
QB*	3.62%	5.72%	28.50%	12.46%	7.37%	2.24%
RB	11.87%	6.22%	0.07%	10.56%	3.96%	0.20%
TE	7.88%	2.76%	5.84%	5.01%	1.62%	0.11%
WR	3.88%	2.47%	2.53%	3.51%	0.47%	0.33%

* = very low sample size of quarterbacks who participated in combine events

Another application of the R^2 output is to compare the importance of the NFL Combine across positional groups. For example, the position of center and offensive guard (C/OG) is significantly less affected by combine performance than the position of offensive tackle (OT). The top three R^2 values for centers and offensive guards are the broad jump, 40-yard dash, and vertical jump, which are 3.59%, 2.92%, and 1.38% respectively; in comparison, the top three R^2 values for offensive tackles are the 40-yard dash, broad jump, and bench press, which are

9.11%, 5.93%, and 5.31% respectively. The differences in these values show that the combine has a larger impact on draft stock for tackles than it does for centers and guards. This analysis remains consistent with common practice in NFL scouting, as the athleticism of tackles is looked at far more heavily than that of centers and guards when projecting value for offensive linemen at the next level. Overall, this analysis can be beneficial for players and their agents when preparing for the NFL Combine.

The Draft Projection Calculator

The primary feature of this data model is the Draft Projection Calculator. This is the main function of the regression formulas generated from the training data of NFL Draft and Combine results. This calculator can be used to tell a player how improvements in each of the six combine events can positively affect his draft stock. The analyst, or whoever may be operating the calculator, must use the drop-down menu to select the position of the player, for example “RB” for a running back, and then insert their performance results for each of the combine events in the column labeled “BEGIN.” In order to test the effect of potential improvements, the operator of the calculator must then insert the performance results they seek to compare to their originals, perhaps the results from intense training or the goals set by the player. An example of this can be seen in the picture below, using 2018 running back Royce Freeman’s combine results in the “BEGIN” column and hypothetical incremental improvements for each event in the “END” column:

Position		RB
Combine Event	**BEGIN**	**END**
40yd	4.54	4.50
Vertical	34	37
Bench	17	20
Broad Jump	118	125
3 Cone	6.90	6.68
Shuttle	4.16	4.07

From here, the calculator will do the rest of the work. I programmed the calculator to reference the regression formulas from the analysis in order to predict where the player will be drafted based on his combine performance. It does this for both the BEGIN and END columns, and then calculates the average projected draft position based on the six events. For example, the formula to predict where a running back will be drafted based on his 40-yard dash time is y = 214.8221(x) - 855.37. Since Royce Freeman ran a 4.54 in the 40-yard dash, his projected draft position is 119.92. After generating the six projected draft positions associated with the six variables in each column, it will calculate the average expected draft position based on the player's combine performance. The result of this process can be seen below:

Draft Pick	**BEGIN**	**END**
40yd	119.92	111.33
Vertical	129.31	113.27
Bench	125.68	124.52
Broad Jump	130.02	102.57
3 Cone	117.78	102.73
Shuttle	122.35	120.49
AVERAGE	124	112

This output can be interpreted to mean that using this model, Royce Freeman's projected draft position would have been 124th overall, but if he was able to make the improvements listed in the "END" column, his projected draft position would improve to 112th overall.

The Application of the Model

It is important to note that the Combine is not the sole determinant of a prospect's draft value since scouts will consider factors such as on-field production, scheme fit, and overall talent in addition to the athleticism measured at the NFL Combine. Because of this, the predicted draft position should be interpreted on a relative basis, instead of as an exact indication of where a player will be selected. For this reason, I believe that the real value of this model lies in the percentage increase in draft stock. Furthermore, this analysis does not mean that Royce Freeman will be selected with the 112th overall pick in the draft if he improves his combine performance. What it really means is that the model predicts that his draft stock will improve by 9.4% if he improves his combine results. This output can be seen in the model as follows:

INCREASE (PICKS)	**11.7**
INCREASE (%)	**9.4%**

Conclusion

I believe that this concept of percentage increase is the true value of this data model. A player can utilize this analysis to determine which events they should focus on in preparation for the combine to most effectively improve their draft stock. They can use the R^2 value chart to

determine which events are the most significant for their position group, and they can use the Draft Pick Calculator to calculate the projected increase in draft stock that they can expect if they improve their combine performance. In a further study, it would be beneficial to quantify the other aspects of prospect evaluation (i.e. production, scheme fit, NCAA conference, etc.) in order to do a full-scale analytical draft prediction. The Draft Pick Calculator was a good place to start, however I would love to do a more in-depth analysis on the entire evaluation process for the NFL Draft in the future.