Machine Learning: Predicting Draft Positions of NFL Prospects from 2009 to 2015

Adam Bruce

2024-04-12

Background

Each year the 32 teams of the National Football League (hereafter NFL) come together to select players from the nations top collegiate prospects to be a part of their team. This meeting, known as the NFL Draft began in 1936, and has since been fine-tuned to span seven rounds. Every team is allocated one selection per round, but they possess the ability to trade and acquire selections from each other. Ultimately, the likelihood of a prospective player succeeding at the professional level diminishes as the rounds progress. As a result, this makes early round picks extremely valuable to the holders. However, what makes a prospect translate into a high round selection isn't exactly clear. This uncertainty leads to teams spending endless financial and human resources scouting for physical and schematic attributes they believe will translate to future success.

Each year in February, two months prior to the actual draft, at the NFL Scouting Combine the peak of the scouting process is reached. The combine gives players a chance to showcase their value through performance events and meetings. Routinely, the event leads to some prospects desirability falling and others rising, which highlights the question of what attributes are most indicative of a player rising or falling. In hopes of addressing this question, we primarily aim to use data on prospects from the 2009 to 2019 renditions of the NFL Scouting Combine to predict whether they were an "Early Round" (rounds 1-3) or "Late Round" selection (rounds 4-7) in the draft. Additionally, we aim to understand which attributes are most influential, and perhaps how so, to our best models predictions.

Data Preparation and Exploration

Table 1: Percent Missing Values Before Removing 698 Missing Rows for Cone Drill and 280 Missing Rows for Bench Press Reps (LEFT) Versus After Removal (RIGHT)
2009 to 2019 NFL Combine Data

Variable	Mean	Total_NAs	Percent_Missing_Before
Age (years)	22.00	11	0.50
Height (inches	74.00	0	0.00
Weight (pounds)	245.00	0	0.00
40 Yard Dash Time (seconds)	4.75	57	2.58
Vertical Jump (inches)	33.17	368	16.64
225 Pound Bench Press Reps	21.00	503	22.74
Broad Jump (inches)	115.56	389	17.59
30 Yard Three Cone Drill Time (seconds)	7.22	698	31.56
20 Yard Shuttle Drill Time (seconds)	4.39	654	29.57
Body Mass Index (kg/m^2)	31.33	0	0.00
Player Type (Offense or Defense)	NA	0	0.00

Variable	Mean	Total_NAs	Percent_Missing_After
Age (years)	22.00	7	0.32
Height (inches	74.00	0	0.00
Weight (pounds)	249.00	0	0.00
40 Yard Dash Time (seconds)	4.76	10	0.45
Vertical Jump (inches)	33.11	10	0.45
225 Pound Bench Press Reps	21.00	0	0.00
Broad Jump (inches)	115.06	21	0.95
30 Yard Three Cone Drill Time (seconds)	7.23	0	0.00
20 Yard Shuttle Drill Time (seconds)	4.40	16	0.72
Body Mass Index (kg/m^2)	31.85	0	0.00
Player Type (Offense or Defense)	NA	0	0.00

Originally, the NFL Combine data had 3,477 observations for 18 variables. Immediately, 1,223 observations were removed because they could not be specified as early or late round selections since they were from players that completed the combine but were not drafted. Next, the variables "Year", "Player", "Position_Type", "Drafted", "School", and

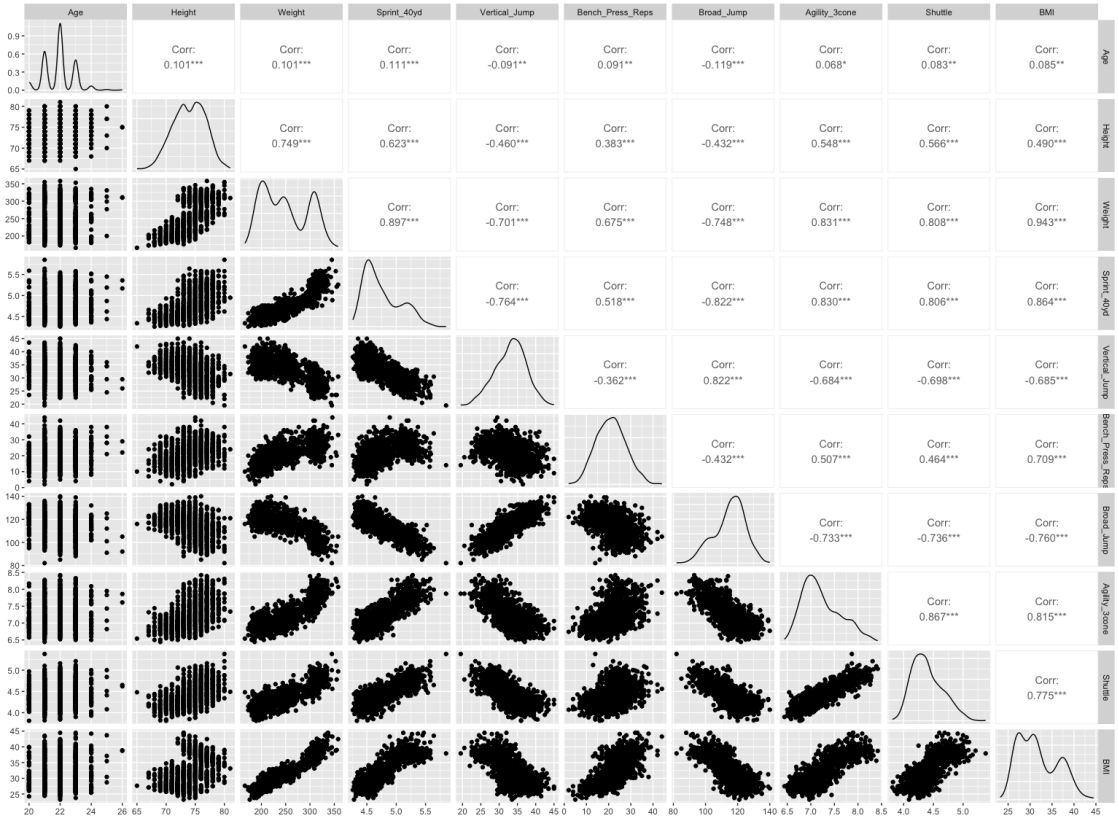
“Position” were all dropped because they were not useful for the goals of our modeling or were not applicable to the models used. To extract the information for our response, the variable “Drafted..tm.rnd.yr.” was split into four unique variables in “Draft_Team”, “Draft_Round”, “Pick_Number”, “Selection_Year”. Of these, only “Draft_Round” was kept. Since we were interested in classifying early versus late round choices, we converted “Draft_Round” values of 1st, 2nd, or 3rd into ‘Early Round’ and values of 4th, 5th, 6th, and 7th into ‘Late Round’ subcategories.

Now, the dataset contained 11 explanatory variables and our response. The meaning of each explanatory variable can be found in Table 1 (Above). However, “Height” was converted from meters to inches, “Weight” from kilograms to pounds, “Vertical_Jump” from centimeters to inches, and “Broad_Jump” from centimeters to inches. After the conversions, how to deal with the extensive missing data in the dataset had to be determined.

Overall, there were 2,932 cases of missing data for the 2,254 remaining observations. A total of 42 athletes were missing values for all six event-based variables, so they were dropped. Next, the means and total number of missing values were calculated for each explanatory variable (Table 1: Above, LEFT). Using these values, we aimed to impute the mean for each variable if and only if 10% or less of the variables observations were missing. However, we found 698 individuals (31.56%) were missing a value for the “Agility_3cone” variable. Therefore, players with missing values for this variable were filtered out. After removal, 1,514 observations remained, with 12% of players missing values for “Bench_Press”. Therefore,these players were also filtered out, leaving 1,234 observations.

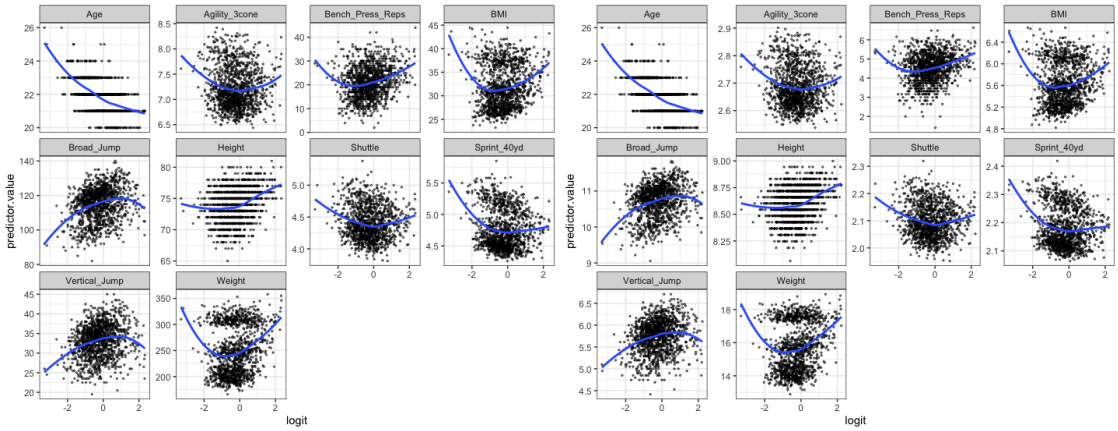
After cleaning, the means and total number of missing values were calculated for each explanatory variable (Table 1: Above, RIGHT). We observed the highest frequency of missing values to be less than 1% (Broad Jump 0.95). Therefore, imputation of the means for each variable with missing values was performed. To ensure a necessary amount of observations for each level of the response, we checked the counts and found 552 athletes were “Early Round” selections while 682 were “Late Round” choices. This was an adequate split for accurate predictive modeling, so we proceeded with our cleaned data for exploratory data analysis.

Figure 1: Assessing Multicolinearity Among Numeric Predictors in the NFL Combine Dataset



In Figure 1 (Above), multicolinearity among independent predictors is assessed for the data. Of the possible two variable combinations, 20 cases achieved a correlation score greater than 0.70 or less than -0.70. This was a clear indication that many predictors were highly correlated, which made RIDGE, LASSO, or ENET logistic regression a good option for the data. However, it was necessary to check for linearity of independent variables and log odds because our response variable was binary (logistic regression based).

Figure 2: Numeric Predictor Variable Values VS Logit for Non-Transformed (LEFT) and Square Root Transformed (RIGHT) NFL Combine Logistic Regression Models



When assessing linearity of independent variables and log odds, we observed a non-linear relationship for the numeric variables (Figure 2: Above, LEFT). It was clear that the relationship more closely resembled a parabolic shape rather than a linear one. Therefore,

we attempted a square root transformation of the all predictors to attempt and address this concern (Figure 2: Above, RIGHT). However, this was unsuccessful in achieving linear relationships. Ultimately, this made it clear that non-linear approaches would be optimal for making predictions with this dataset. Knowing the modeling approaches would have to be robust to multicolinearity, we moved forward with Random Forest and Artificial Neural Network techniques. To do so, we had to use one-hot encoding for the variable "Player_Type" where 0 represented defense and 1 represented offense for player "Position_Type". This was necessary because Artificial Neural Networks cannot generalize a "best" function with categorical data.

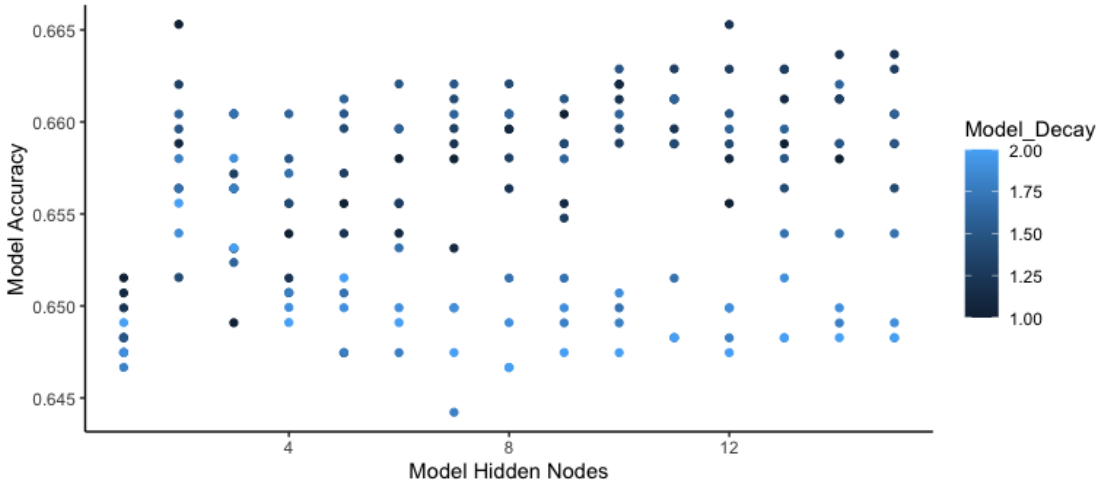
Model Fitting

Single Cross Validation

Table 2: Results of Single 10-Fold Cross Validation with Artificial Neural Network and Random Forest for the NFL Combine Dataset

Model	Parameter_Range	Optimal_Parameter	Maximum_Accuracy
Artificial Neural Network	Size = 1:15 and Decay = seq(1, 2, by = 0.1)	Size = 2 and Decay = 1	0.6653
Random Forest	mtry = 1:11	mtry = 3	0.6378

Figure 3: Highest Accuracy is obtained with 2 Hidden Nodes and a Weight Decay of 1 for 10-Fold CV with an Artificial Neural Network on the NFL Combine Dataset



Utilizing a single round of 10-Fold Cross Validation on the NFL Combine dataset, an Artificial Neural Network and Random Forest model that optimizes accuracy of predictions was obtained. Overall, Table 1 (Above, Top) shows the range of parameters used in cross validation for each model respectively. It is important to note that our Artificial Neural Network model did converge, meaning either a local or global minimum was found for the gradient function. Additionally, we used a range of hidden nodes between 1 and 15 because it encompassed a range slightly greater than our number of predictors plus the levels of our response variable. As for decay, we went with a penalty sequence between 1 and 2 in order to not risk the model becoming overly bias with shrinkage.

Overall, an Artificial Neural Network model with parameters for hidden nodes (size) and weight decay (decay) considered at each split was optimal during this cross validation process. This model obtained a maximum accuracy of 0.6653 (66.53%) when predicting "Draft_Round" while the best Random Forest model obtained a lower maximum accuracy of 0.6378 (63.78%). Figure 3 (Above, Bottom) showcases the accuracy values obtained at each number of Hidden Nodes and Decay for the cross validation process. Ultimately, the maximum accuracy at 2 hidden nodes with a decay value of 1. Going forward, we will perform double cross validation with a 5-fold outer split and 10-fold inner split to honestly assess our model's prediction performance on new, unseen data using a train/test split. Then, we will assess our "best" model obtained here.

Double Cross Validation

Table 3: Results of Double 5-Fold Cross Validation with Artificial Neural Network and Random Forest Models for the NFL Combine Dataset

Outer_Loop	Optimal_Model	Optimal_Tune	Maximum_Accuracy
1	Artificial Neural Network	Size = 2 & Decay = 1	0.6503
2	Artificial Neural Network	Size = 1 & Decay = 1.6	0.6678
3	Artificial Neural Network	Size = 5 & Decay = 1.2	0.6617
4	Artificial Neural Network	Size = 2 & Decay = 1.2	0.6697
5	Artificial Neural Network	Size = 13 & Decay = 1.5	0.6701

Figure 4: NFL Combine Confusion Matrix for Honest Predictions Assessment with 5-Fold Double CV

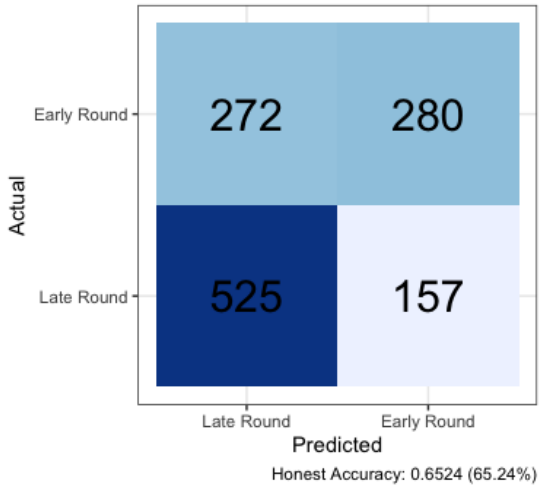
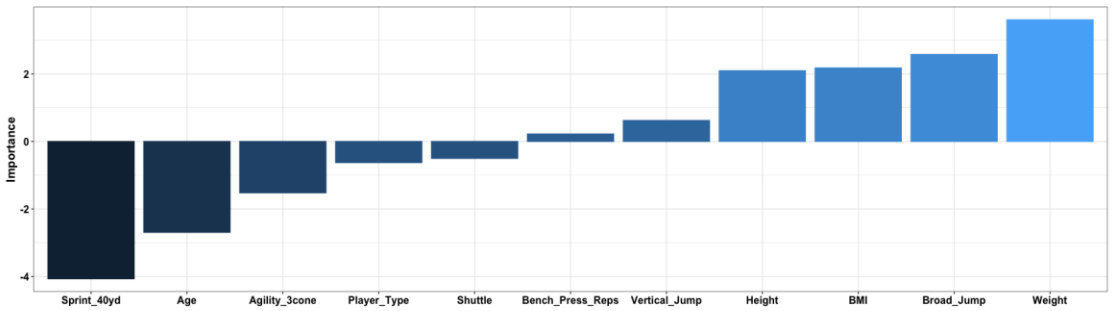


Table 3 (Above, Top) provides the output for the optimal models at each 5 outer CV fold. Overall, Artificial Neural Networks always optimized accuracy better than Random Forests. For accuracy, we are most concerned with predictions on the test/validation set at each split. At each fold, we fit the optimal model to the test/validation set to obtain predictions for the set observations. These predictions were then saved before moving onto the next fold.

Utilizing the confusion matrix in Figure 4 (Above, Bottom), we observe how the optimal models in Double CV performed in their predictions. However, most importantly, the honest accuracy can be calculated at 805 correct predictions out of 1234 observations. We see clearly that the model performs much better at predicting late round selections (525 of 682 or 76.98%) compared to early round selections (280 of 552 or 50.72%). Overall, our honest accuracy was 0.6524 (65.24%), which indicates we can expect our "best" model from Single CV to predict the Draft Round of an NFL Combine prospect correctly on average about 65.24% of the time.

Model Interpretation

Figure 5: Variable Importance Via Oldens Algorithm of the Best 10-Fold Single Cross Validation Artificial Neural Network for the NFL Combine Dataset (ABOVE) with Absolute Value Variable Importance Table (BELOW)



Variable	Importance
Sprint_40yd	4.0679
Weight	3.6042
Age	2.6924
Broad_Jump	2.5771
BMI	2.1774
Height	2.0958
Agility_3cone	1.5221
Age	0.6267
Vertical_Jump	0.6224
Shuttle	0.5024
Bench_Press_Reps	0.2212

Using our best Artificial Neural Network model with Size = 2 and Decay = 1 from 10-fold single cross validation, a variable importance plot was constructed via Olden's Algorithm (Figure 5: Above, TOP). This plot provides the opportunity for assessing not only the importance for each variable in predicting "Draft_Round", but also the relationship of the variable with the response "Early Round". Variables with positive importance measures are positively associated with predicted probabilities of "Early Round" selections, and those with negative importance measures are negatively associated with predicted probabilities of "Early Round" selections. However, the largest bars in either directions are the most influential, so an absolute value table helps with understanding overall importance (Figure 5: Above, BOTTOM).

Ultimately, we observe that the most important variable was "Sprint_40yd", which was negatively associated with the predicted probability of an "Early Round" selection. This makes sense, as speed is an essential part to every position in the NFL. In fact, some teams are even known for drafting players highly solely based on their speed. One can find numerous drafts where analysts harp on teams for drafting players too early solely based on their speed. Overall, speed is viewed as critical, but athletic traits are also important, which is why these choices have been so scrutinized.

The second most important variable to the model was weight, which is positively associated with the predicted probability of an “Early Round” selection. This also makes sense as American Football is a violent contact sport where athletes with large physical statures tend to dominate. Clearly, a 260 pound defensive player taking on a 160 pound offensive player on the field would not favor the offensive player in most cases. Thus, having an above average weight for your respective position would make a player enticing for any team.

Finally, to touch on the 3rd and 4th most important variables, “Age” is negatively associated with the predicted probability of “Early Round” selections. The most fitting explanation for this comes from the saying, “NFL or Not-For-Long.” The vicious nature of the game catches up to players fast. In fact, the average career of an NFL player is only 3.3 years. Therefore, teams almost always prefer to draft younger players as opposed to older ones during the draft. As for Broad Jump, this is a test seen by most analysts and coaches as a major indicator of athleticism. Sometimes, the most talented college players are not the most successful when they enter the NFL because they lack athleticism. Thus, teams covet great athletes, which explains why “Broad_Jump” is positively associated with the “Early Round” response.

To further illustrate and finalize the relationships for the top four variables in the final Artificial Neural Network Model, Figure’s 6 through 9 (Below) provide partial dependence plots for the predicted probabilities of an “Early Round” or “Late Round” model response. Note that these plots assume all other variables are held constant at their medians.

Figure 6: Partial Dependencies of Weight for the Draft Round NFL Combine Response Variable

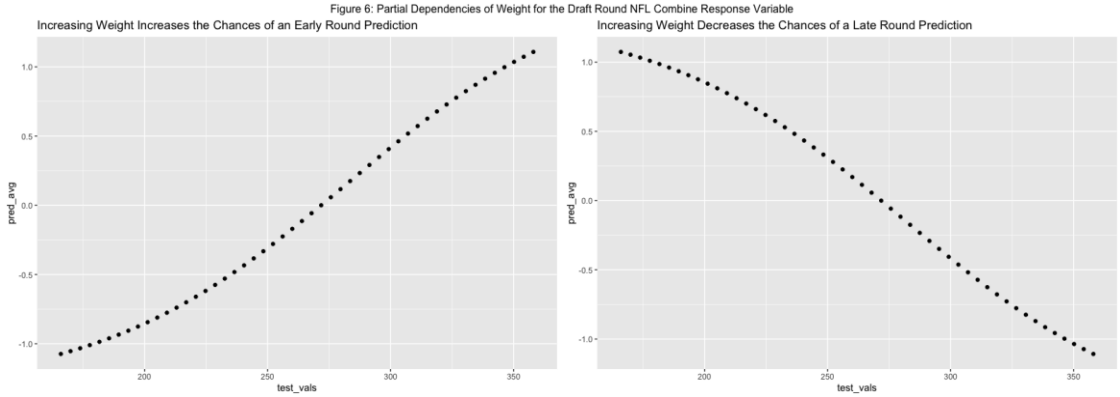


Figure 7: Partial Dependencies of Broad Jump for the Draft Round NFL Combine Response Variable

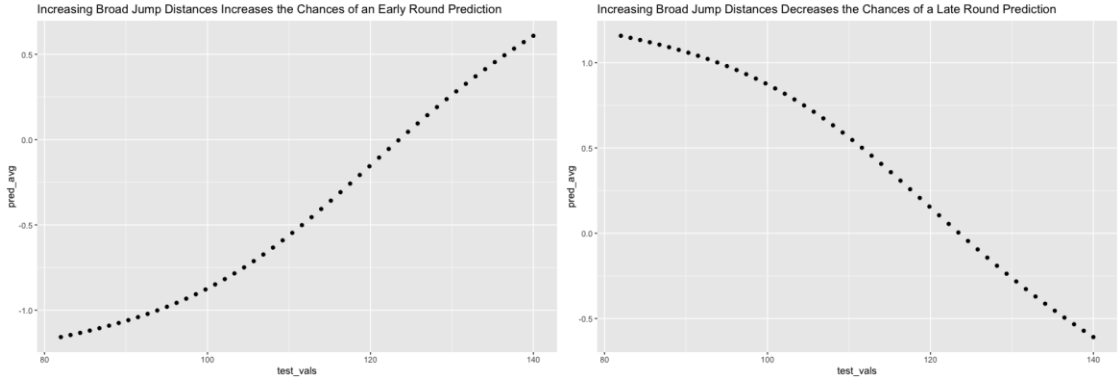


Figure 8: Partial Dependencies of 40 Yard Dash Times (Sprint_40yd) for the Draft Round NFL Combine Response Variable

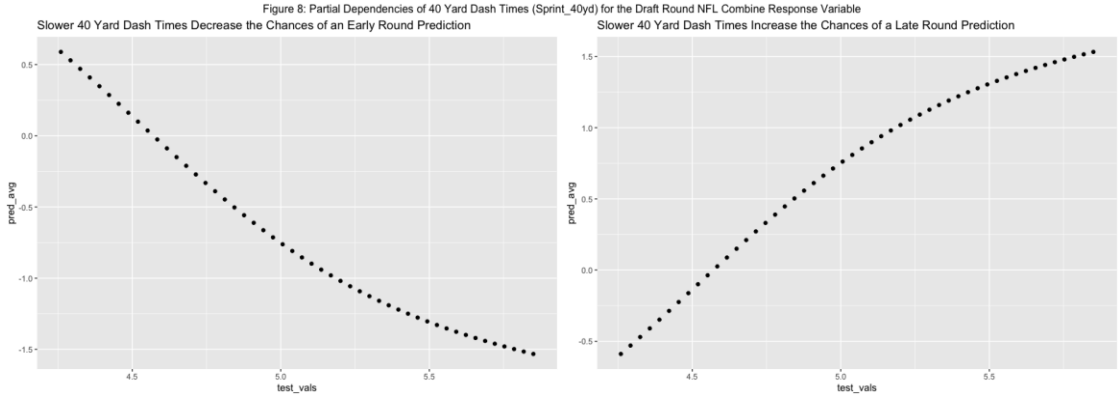
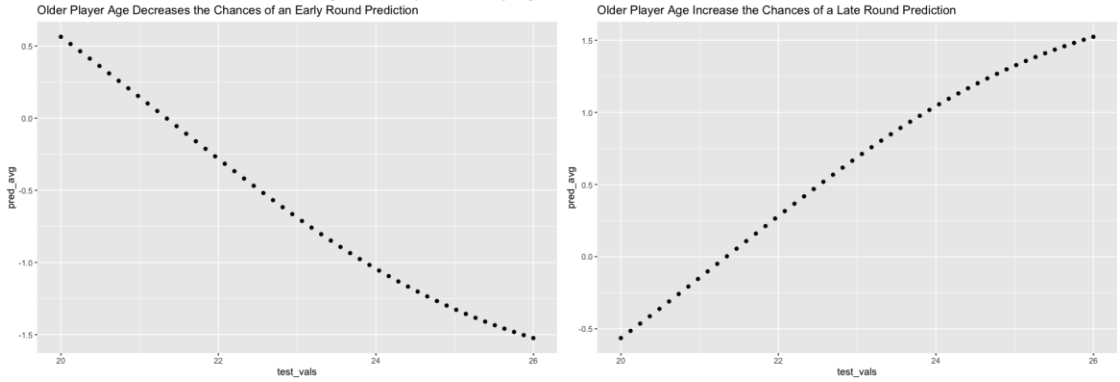


Figure 9: Partial Dependencies of Player Age for the Draft Round NFL Combine Response Variable



Conclusions

Overall, we observed an honest model accuracy of 65.24% based on assessment with 5-Fold Double CV. Using our confusion matrix, we see that the Draft Round category of "Late Round" selection has the highest frequency of observations in the dataset at 682. By taking this value divided by the total number of observations, 1234, we receive a No-Information Rate of: 0.5527 (55.27%). Thus, if we were to randomly assign any new observation to be the majority class "Late Round", we would predict the class correctly about 55.27 percent of the time. This means that, on average, our model will predict the Draft Round of an NFL Combine prospect 9.97% more often than randomly guessing. This provides some level of support for the final Artificial Neural Network model being sufficiently accurate to be used as a prediction model going forward.

Ultimately, the context of how we view our final model needs to be subjected to its use. If we were predicting Heart Diseases in patients, then we would probably not be satisfied with only a 9.97% increase in prediction accuracy compared to the No-Information Rate. However, our model looks to assess a sport at its highest level of competition. Therefore, the talent gap at this level is extremely minuscule, and an approximately 10% better chance of achieving a desired outcome for an NFL team would be very desirable. Thus, we would say the 9.97% increase and 65.24% accuracy in correct predictions on average is sufficient to use the model going forward.

However, though we view our model as sufficient, there are certainly other useful predictors for "Draft_Round" of NFL prospects not accounted for in this assessment. Two

such variables would be Football Intelligence Quotient (FIQ) and Off-Field Issues. A high FIQ is extremely sought after in the NFL. A player's ability to read the game correctly during each play can be the difference between a win and a loss. There are several tests to quantify this measure, and including it for our model would almost certainly help increase our accuracy. Finally, teams are always turning away from prospects that get into legal trouble away from the game. A player may have all the attributes desired for success, but that means nothing if they cannot even suit up for games. Gathering additional variables such as these could be useful in producing an even better prediction model in the future.