Michael Pallante
MSDS 456
Assignment 1

Every year, the National Football League (NFL) hosts the NFL Scouting Combine event, where over 300 prospects have an opportunity to showcase their physical abilities in a series of different drills. This event occurs about a month before the NFL Draft, and teams get a chance to observe prospects in a professional setting and get a look at their work ethic. Typically, the NFL Combine can lend scouts a view of prospects' athletic capabilities and other physical measurables, including size, speed, strength, and agility. The event is aired on national television for a week and gains a lot of viewership. Everyone that pays attention to the NFL nationally has their opinion on how much the NFL Combine event truly matters to a player's draft status leading up to the NFL Draft. Through this project, we will be taking a further look into that topic. The problem we will be exploring is how much the NFL Combine affects a player's draft stock leading up to the NFL Draft. We will be using a data set consisting of the last five years of the NFL Combine, from 2013 to 2017.

Before we can enter the modeling phase, we must prepare the data set for further use. We will be dividing the data into a training data set and a test data set to try to improve the accuracy of our modeling results. The training data set will be composed of all NFL Combine data from the years 2013 to 2016, and the test data set will only contain the data from the year of 2017. We will be building our models on the training data set and testing them on the test data set. We will also be developing a null model to use as a comparative measure to judge our model against. The predictive performance of these models will be evaluated by the computation of their root mean-square error (RMSE) value, where the lower the RMSE value, the better fit and better performing the model is. Next, we must take a look at all of the variables at our disposal and understand the data we are using. In Figure 1.0 below, we can see an overview of the type of data on hand.

Figure 1.0: Overview of NFL Combine Training Data Set

```
> str(training)
'data.frame':	846 obs. of  18 variables:
 $ Player     : chr  "A.J. Klein" "Aaron Dobson" "Aaron Mellette"...
 $ Year       : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
 $ Rnd        : int  5 2 7 4 4 6 1 4 6 7 ...
 $ Pick       : int  148 59 238 101 100 195 30 103 187 217 ...
 $ Tm         : Factor w/ 32 levels "ARI","ATL","BAL",..: 5 21 3 15 30 13 ...
 $ ID         : chr  "KleiAJ00" "DobsAa00" "MellAa00" "SandAc00" ...
 $ PosPro     : chr  "LB" "WR" "WR" "WR" ...
 $ Age        : int  22 22 23 21 21 22 21 22 24 23 ...
 $ CollegeUniv: chr  "Iowa St." "Marshall" "Elon" "South Carolina" ...
 $ PosCollege : chr  "ILB" "WR" "WR" "WR" ...
 $ Ht         : int  73 75 74 67 73 70 74 76 69 76 ...
 $ Wt         : int  250 210 217 173 307 193 242 264 199 263 ...
 $ X40yd      : num  4.66 4.37 4.54 4.58 5.15 4.59 4.7 4.78 4.61 4.86 ...
 $ Vertical   : num  NA NA 33.5 32 30 33 33.5 NA 34 31.5 ...
 $ Bench      : int  20 NA 9 7 37 14 20 21 NA NA ...
 $ BroadJump  : int  113 NA 123 117 107 117 122 NA 122 118 ...
 $ X3Cone     : num  NA NA 7.11 6.81 7.82 NA 7.16 NA NA 7.32 ...
 $ Shuttle    : num  NA NA 4.41 4.37 4.72 4.15 4.39 NA NA 5 ...
```

The main variables we will be focusing on are our response variable, Pick (the draft pick a player is selected on), and our explanatory variables, which will be composed of all physical and performance-related variables. The physical explanatory variables specifically include height (Ht) and weight (WT), while the performance explanatory variables include the six Combine drills included in the data: 40-yard dash (X40yd), Vertical Jump (Vertical), Bench Press (Bench), Broad Jump (BroadJump), Three Cone Drill (X3Cone), and 20 Yard Shuttle (Shuttle). The model we will be comparing to the null model will be referenced as the evaluation model (eval model), which will be composed of a series of different models (component models) combined into one large model (the eval model) in which we will compute the average RMSE through cross validation techniques and then weigh it against the average RMSE of the null model. The methodology for our component models that was deemed best was creating multiple linear regression models (lm) by player position, using the three most important performance tests for each position, plus their physical traits (Ht and Wt). This will give us unique models for almost every position. Some positions will be grouped together due to having similar qualities and sharing the same three most important performance tests.

Now that we have established what we will do in the modeling phase, we must perform a missing data test on our data set to check for null observations before we can proceed to the modeling. The missing data test can be seen in Figure 2.0 below, where each variable is accompanied by the percentage of null observations of that variable.

Figure 2.0: Missing Data Test

Variable	Percentage of Null Observations
X3Cone	33.5697400%
Shuttle	30.8510638%
Bench	25.6501182%
BroadJump	18.7943262%
Vertical	18.3215130%
X40yd	1.0638298%
Age	0.1182033%
Player	0%
Year	0%
Rnd	0%
Pick	0%
Tm	0%
ID	0%
PosPro	0%
CollegeUniv	0%
PosCollege	0%
Ht	0%
Wt	0%

As we can see, there is a large percentage of null observations within all six of our NFL Combine performance test variables. This is not abnormal, as many players choose not to participate in all Combine drills, and some players even choose not to attend the Combine event at all. So, this missing data was to be expected. It is such a large number of missing data that we have to be careful with the imputation technique that we select, because we do not want to use any fake or made up data and we do not want to create any bias in our analysis. To offset the big presence of missing data here and not having any influence on our data set, we will be creating an ensemble or hybrid prediction in which we average all of our component model predictions with non-missing values only, and one player at a time.

Next, we must identify each of our component models that will be featured in our evaluation model, and also explain how we identified the specific Combine tests that were viewed as most important to each position. In Figure 3.0 below is a brief explanation of each model that was built for each position and justifies our choices of explanatory variables for these positions. Keep in mind that we have already mentioned that each model will feature Height (Ht) and Weight (Wt) because we want to know how much size impacts a player's draft status.

Figure 3.0: Description of Component Models

1. **qb.k.p.model <- lm(Pick ~ Ht + Wt + Shuttle + Vertical + BroadJump)**

   The Quarterback (QB), Kicker (K), and Punter (P) positions were grouped together for this model. The Shuttle drill can probably show us the footwork of a QB and we can see how quick he is on his feet, while we can also see the movement speed and reaction time of a K and P on special teams units. Both the Vertical Jump and Broad Jump can indicate the lower body explosiveness of these positions. For a QB, we will see the power in which he steps into his throws or how firm he can stand in the face of pressure before going down. For a K or P, we can see the type of leg power they have and if it can translate into raw kicking power.

2. **rb.model <- lm(Pick ~ Ht + Wt + X40yd + X3Cone + BroadJump)**

   This model is exclusively for the Running Back (RB) position. The 40 Yard Dash is crucial for a running back in that we can determine his breakaway speed, which can be the difference between his ability to get a modest gain on the ground or create a game changing play. We can also see if he has long speed as opposed to just quickness. The 3 Cone Drill is important in assessing a RB's acceleration, agility, and quickness. It may also give us an indication of a RB's elusiveness in the open field. The Broad Jump is a good view of a RB's lower body explosiveness. This may tell us if he is able to generate power from the ground up to break tackles or hold down a block in pass protection against a larger defensive player.

3. **wr.model <- lm(Pick ~ Ht + Wt + X40yd + Vertical + X3Cone)**

   This model is exclusively for the Wide Receiver (WR) position. Similarly to the RB position, the 40 Yard Dash is crucial for a wide receiver in that we can determine his breakaway speed, which can be the difference between his ability to get a modest gain after the catch or create a game changing play. It can also be a huge advantage in the ability to create separation from cornerbacks. The Vertical Jump is important in assessing a WR's ability to get up for a jump ball or catch balls out of the air in traffic. The 3 Cone Drill is important in assessing the acceleration, agility and quickness of a WR. It may also give an indication of how smoot a WR is with his feet and may suggest if he can get in and out of his routes quickly.

4. **fb.te.ls.model <- lm(Pick ~ Ht + Wt + Bench + Vertical + BroadJump)**

The Fullback (FB), Tight End (TE), and Long Snapper (LS) positions were grouped together for this model. The Bench Press can show the upper body strength and raw strength of any of these positions to hold blocks. Both the Vertical Jump and Broad Jump can show the lower body explosiveness of these players off the line of scrimmage and their ability to create power from the ground up, which is also key in gaining leverage in the blocking game, or FB's or TE's being able to break a tackle with the ball in their hands. It can also show a FB or TE's explosiveness in their route running in the passing game.

5. **ot.model <- lm(Pick ~ Ht + Wt + Shuttle + X40yd + Bench)**

   This model is exclusively for the Offensive Tackle (OT) position. The Shuttle can show an OT's reaction time and their ability to not only pick up their blocks at the line of scrimmage, but also gain leverage on pass rushers. The 40 Yard Dash can be used to show an OT's initial burst off the line of scrimmage, indicated by the 10 yard splits that are timed in the 40 Yard Dash, and it can also show an OT's ability to get downfield to make blocks on the second level of the defense. The Bench Press can show the upper body strength and raw strength of an OT to hold blocks.

6. **og.c.dt.model <- lm(Pick ~ Ht + Wt + Shuttle + Bench + BroadJump)**

   The Offensive Guard (OG), Center (C), and Defensive Tackle (DT) positions were grouped together for this model. The Shuttle can show the reaction time of an OG or C and their ability to not only pick up their blocks at the line of scrimmage, but also gain leverage on pass rushers. The Bench Press can show the upper body strength and raw strength of an OG or C to hold blocks, or a DT's ability to control and stack blockers and then shed their blocks (the football term is stack-and-shed). The Broad Jump can show the lower body explosiveness of these players and their ability to create power from the ground up, which is also key in gaining leverage in the blocking game for an OG or C. It also shows the burst of a DT to break through the line of scrimmage and blocks to be able to create a pass rush. Additionally, it also shows the athleticism of an interior offensive lineman to be able to swing to the OT position if needed.

7. **de.prolb.model <- lm(Pick ~ Ht + Wt + X3Cone + Vertical + BroadJump)**

   The Defensive End (DE) and pass rushing Outside Linebacker (prOLB) positions were grouped together for this model. I decided to classify OLB into two separate models but did not change any of their data to make that possible and was just demonstrating the skills they'd need to be successful in different capacities. The 3 Cone Drill demonstrates the ability of these positions to change direction quickly, bend and accelerate, and edge rush. The Vertical Jump and Broad Jump can both show the lower body explosiveness of these players and their ability to create power from the ground up to break through or run past the line of scrimmage and blocks to be able to create a pass rush.

8. **ilb.tolb.model <- lm(Pick ~ Ht + Wt + Shuttle + X3Cone + BroadJump)**

   The Inside Linebacker (ILB) and traditional Outside Linebacker (tOLB) positions were grouped together for this model. As referenced in the previous model, I decided to classify OLB into two separate models but did not change any of their data to make that possible and was just demonstrating the skills they'd need to be successful in different capacities. The Shuttle shows these positions' reaction times and their ability to quickly diagnose a play on defense. Along with the 3 Cone Drill, it also shows these players' ability to stop and explode toward the ball and the body control needed to change directions quickly. The Broad Jump can show the lower body explosiveness of these players and their ability to create power from the ground up to shed blocks to make tackles in the open field and also may show if the players can launch to make strong tackles.

9. **cb.s.model <- lm(Pick ~ Ht + Wt + Shuttle + X40yd + X3Cone)**

   The Cornerback (CB) and Safety (S) positions were grouped together for this model. The Shuttle shows these positions' reaction times and their ability to quickly diagnose a play on defense. Along with the 3 Cone Drill, it also shows these players' ability to stop and explode toward the ball and the body control needed to change directions quickly. These drills also shows these players' ability to change directions to keep up with the routes of opposing WR's and TE's in the open field. The 40 Yard Dash shows the CB's ability to keep up with WR's downfield in coverage and shows the S ability to cover a lot of ground in the deep middle of the field in addition to coverage responsibilities when needed.

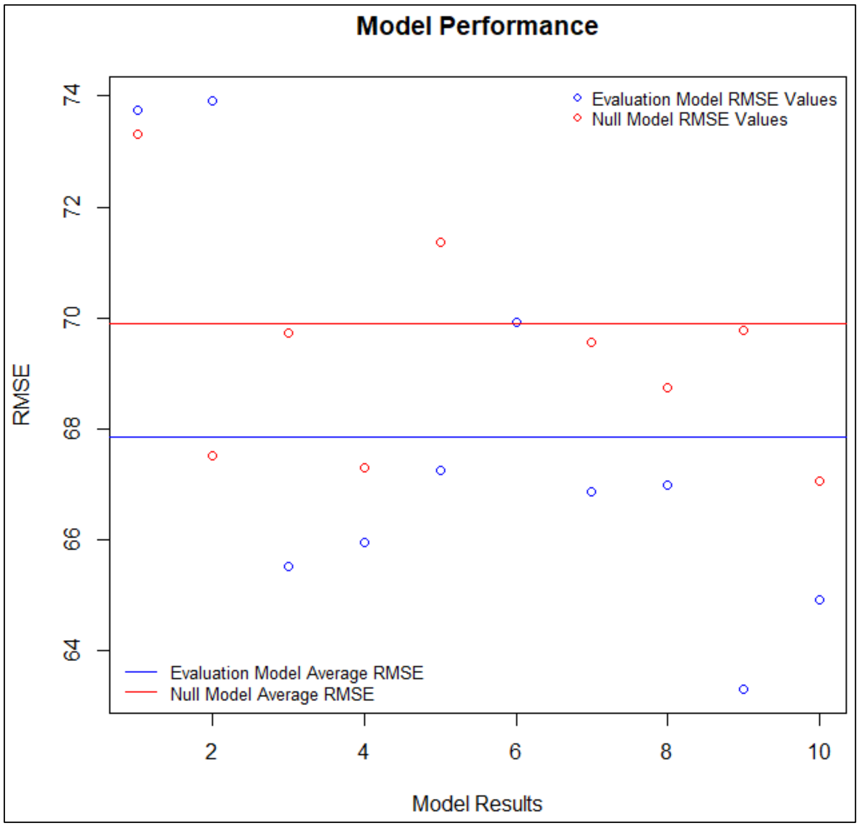
After we identified the component models that compose the evaluation model, we can now run the evaluation model. Remember, this model was run on the training data set. After running the evaluation model, we then predict the entire evaluation model and its component models on the test data set. Once that is completed and cross validation techniques are complete, we are able to judge the performance of our evaluation model against the null model. Again, the performance metric for these models is the computation of their average root mean-squared error (RMSE) values, and the lower the RMSE the better fit and better performing the model is. In Figure 4.0 below, we can see the performance of both models in comparison to each other.

Figure 4.0: Model Performance

Model	Average RMSE Value
Evaluation Model (Our Model)	67.84188
Null Model	69.90561

The good news is that our evaluation model did outperform the null model and had a lower average RMSE value, proving to be a better fit. This was not done by a large margin though. This is likely due to number of reasons, which we will identify shortly. In Figure 5.0 below, we can see the performance of both models on a plot together, which can help us visualize the results better.

Figure 5.0: Model Performance Plot



This assignment presented us with a number of challenges that likely impacted our results. There were small sample sizes of player data and a large amount of missing data due to players not attending the Combine event or not participating in all drills. Multicollinearity existed among some of our measures. Nonlinear relationships existed between some explanatory variables and the response variable. The importance and relevance of some evaluative measures varied by player position. The quality of players and their test performance varied each year, causing measurement instability over time. There was also a lack of measures for other important attributes of these players. Lastly, physique and NFL Combine performance are not the only evaluative measures of players going into the NFL Draft, and in fact are a very small piece of the pie.

Many of these challenges can be addressed to a degree with the following management recommendations. Increasing the collection of data could improve our chances of capturing player performance in the NFL Combine. the lack of data leaves a lot to be desired when we are trying to measure the true impact of the Combine on a player's draft status. It is hard to account for players that do not participate in the Combine or do not attend at all, however, with a larger sample size of data, we may be able to offset that a little more. The use of data transformations, such as log transformations, can address our multicollinearity issue. Using a Box-Cox transformation can improve the nonlinear relationships between our explanatory variables and response variable. The development of a number of separate models, which we had done, can address the varying value and importance of some of the Combine events to some positions more-so than others. We should not use any models that are not representative of an event's importance to certain positions. For example, a model including Bench Press would not be relevant for a QB, however it would be useful in evaluating the upper body strength of a DT and his ability to stack interior offensive linemen and shed their blocks. So, it would be wise to include the DT position in that model and not include the QB position. Lastly, the NFL Combine is such a minimal part of the evaluation of a player's draft stock, and also their ability to be a productive player in the NFL. More measures are needed to be able to correctly evaluate a player, including game statistics, as well as non-statistical measures including evaluation of game film and evaluation of a player both mentally and football-intelligence-wise. These measures are an incredibly important part of evaluating players too. The inclusion of some of these measures would likely improve our chances of seeing a greater correlation to draft stock than the measures we used in this project. This project simply does not have a sufficient amount of data or evaluative measures to accurately predict a players' draft stock.