Dear Shareholders

As per your request, a study has been conducted into whether or not the draft position of a NFL players can be predicted based on their combine data, which is primarily tests of physical durability, strength and mental agility.

Data was downloaded from the website kaggle () and was distributed to two teams, one using Python and one using R, each assigned to solve the problem in their own way, after which the results were compared.

Based on the result each team concluded that the draft position c*ould not be predicted from combine data with satisfactory precision.*

The main reason being that some entries are missing in the data, and dropping missing data leads to problems with some positions having no or very few observations, which in turn, means that keeping those positions leads to statistical problems, and dropping them means the remaining predictions are both unsatisfactory, since not all positions can be predicted, and correlations between data entries and positions may be skewed due to missing positions.

There's also a general bias in the dataset. The analysis looks at whether or not a draft position could be predicted by a players physical measurements. However a lot of positions are quite similar in their physical needs, like opposing positions such as Wide receiver and cornerback, which both required a high speed, agility and jump characteristics. Hence the model will have a failure at predicting the difference between these. Teams might also have different strategies regarding physical ability and positions, one team might prefer a big player at a position where small players usually reside.

Details of each team will be discussed below.

# Shareholders Paper

In this paper, a short recap of the assignment will be made.

## NFL

The American football league NFL uses a draft system once a year, in order to draft new upcoming pro-players into their teams. We have been working on finding patterns between the newly drafted players physical stats and what position the players are drafted to play. The dataset behind the investigation was found via Kaggle.com.

General for the data:
The data can be downloaded at https://github.com/bande15/sds/raw/master/combine_data.csv
It contains 6219 rows and 16 columns. The columns are distributed as:

1. *Player* - Name of the player.
2. *Position* - The position the player plays at the field, where:
   C = Center, Offensive position, snapper bolden til QB, lineman.
   CB = Cornerback, defensive position, dækker WR op, løber.
   DB = Defensive backs, defensive position, bag lineman.
   DE = Defensive End, defensive position, enden af line overfor TE.
   DT = Defensive tackle, defensive position, lineman ved siden af C.
   EDGE = Det samme som DE.
   FB = Full back, offensive position, beskytter QB.
   FS = Free Safety, defensive position, bag lineman, beskytter zone, midt/stor fyr.
   G = Guard, offensive position, lineman, på hver side af Center.
   ILB = inside linebacker, defensive position, står mellem safety og lineman.
   K = Kicker, special team, sparker bolden i starten af play.
   LB = Linebacker, defensive position, lige bag lineman.
   LS = Underkategori af G, offensive position, lineman.
   NT = Nose tackle, defensive position, det samme som DT, lineman.
   OG = Offensive Guard, offensive position, Overfor Guard.
   OL = Offensive linebacker, offensive position, overfor Linebacker.
   OLB = Outside linebacker, offensive position, overfor FS/SS.
   OT = Offensive tackle, offensive position, overfor DT.
   P = Punter, offensive position, sparker bolden midt i play.
   QB = Quarterback, offensive position, kontrollerer spillet/kaster bolden.
   RB = Running back, offensive position, løber med bolden.
   S = Safety, defensive position, mellem FS og SS, overfor OLB.
   SS = Strong safety, defensive position, anden side end FS.
   TE = Tight end, offensive position, overfor DE, løber efter/med bolden.
   WR = Wide receiver, offensive position, modtager kastede bolde.
3. *Height* - Height of the player in inch.
4. *Weight* - Weight of the player in lbs.

5. *Forty* - How fast the player could run a 40 yard dash measured in seconds.
6. *Vertical* - How high the player can jump vertical measured in inches.
7. *BenchReps* - The number of how many benchreps at 225 lbs the player could take.
8. *BroadJump* - How long a player can jump vertical given in inches.
9. *Cone* - How many seconds a player uses doing a special combined exercise.
10. *Shuttle* - How many seconds a player uses doing a special combined exercise.
11. *Year* - The Year that the player gets drafted.
12. *Pfr_ID* - Player ID.
13. *AV* - All around value.
14. *Team* - The team that draft the player.
15. *Round* - Which round the player was picked .
16. *Pick* - Which number they were picked as.

# Python

Observations with missing values has been excluded from further analysis. Unfortunately this more than halves the number of observations in the dataframe. This also implies that certain positions have very few observations remaining, which in turn means that our model's predictiveness is less accurate.

With the above issue in mind, we continue making an analysis on the dataset.

First off, we have been looking at some explorative data analysis. Here the players were grouped by what team they were drafted to play for, and then the average of the different variables were taken. Doing so showed that the teams in general had similar player compositions, as the average values in the different physical stats were very similar. This might imply, that the teams generally follow a similar strategy when picking players for their teams.

This can be further examined, by doing a similar analysis, grouping players by which position they were drafted to play. Doing so very clearly shows that players in different positions have different physical stats.

Thus we make the hypothesis:

**Player positions implies different skills, thus requiring different physical attributes.**

Focusing on above hypothesis, we can now try to use other techniques to further explore the data, and *maybe* even be able to predict what position a player will be drafted to play, just by looking at their physical stats.

Thus we want to use unsupervised and supervised machine learning to further investigate the dataset.

## Unsupervised learning

The goal of using unsupervised learning, is to look at underlying features in the data. Thus we can ask the question: can underlying factors group the data, and if so: does it group by positions?

Following that question, we made an analysis using different algorithms, with the result being: yes, it is possible, but we do not find enough evidence to conclude, that it is capable of grouping the players by their position.

## Supervised learning

The goal of using supervised learning, is to look at the observations, and try to predict which position they play, based on the other variables. In this analysis, we used five different models, and found that the best model was capable of predicting more than 50 percent correct, with the remaining wrong-predicted positions primarily being grouped into a few other positions. Thus the model can, to a certain degree, predict what a player will be drafted to play, just by looking at their physical stats.

## Python conclusion

Keeping in mind, that a lot of observations got excluded, it still looks like it is partially possible to predict a player's drafted position based on their physical stats. Looking at the results from supervised learning, it also looks like certain positions have very similar physical stats, thus making it difficult for the model to accurately predict the positions. For further investigation, it might be of interest to merge certain positions together.

# R

The purpose of this analysis is to try to find a connection between a player's physical characteristics and the players position. In American football, there are 25 different positions with their own notation. This analysis will be based on a dataset on combine data (The physical evaluation before a player is recruited). The way the R analysis differs from the Python analysis is the way we approach the dataset. In the R analysis there has been made 3 different Unsupervised and supervised analysis, based on groupings of the “Position” variable of the dataset. Some positions in the NFL are quite similar (Center and Guard, Wide receiver and Cornerback, Tight end and Defensive End etc.). Here we had the general thought, that if we grouped some of these under the same factor level, we might have both more observations, but also a better model, since for example Long snapper and Center is the same position, just doing 2 different things.

# 1st analysis

The first analysis looks at the data, with some positions subsetted due to a low amount of observations, such as Quarterback and Long snapper, after NA’s had been dropped. An unsupervised analysis has been made, to look at groupings and characteristics of the dataset. Here we quickly found that positions such as wide receivers, cornerbacks and such were quick with high agility, where players such as Center, Offensive tackle and offensive guard was quite the opposite, with big and sturdy positions.

The supervised analysis were made with an 90/10 split in the data, this is due to the intention of having the most accurate model possible, with a relative low need of verification. The dataset was made with 2-fold crossvalidation, and there were 3 different models predicted from it. A K-nearest model, a decision tree and a random forest. These results are shown in a confusionmatrix to show which errortypes where made. As stated in the introduction, some bias is included in the test, in that a wrong result, might not be a wrong conclusion, as it could just be a player not playing to his intended position as per the model. As expected, most errors were located in and around the same positions, cornerbacks being predicted to be wide receivers and reverse. The results is as following:

| Fit-type | K-nearest | Decision tree | Random forest |
|---|---|---|---|
| Accuracy | 45.16% | 52.33% | 55.56 |

Here we see the most accurate model is the Random forest.

# 2nd analysis

Second analysis is made from a general curiosity of groupings and underlying groups in the dataset. Here we aggregated some positions, and included subgroups into said position. These are EDGE which is a subgroup of DE, NT which is a subgroup of DT, OLB and ILB which is a subgroup of OL, FS and SS, which is a subgroup of S and finally NS which is a subgroup of G.

The same unsupervised and supervised analysis were made for the new dataset, which had the following result in it's prediction algorithm:

| Fit-type | K-nearest | Decision tree | Random forest |
|---|---|---|---|
| Accuracy | 57.95% | 57.95% | 60.78% |

Again the most accurate model is Random forest, though is 2nd analysis is significantly more accurate than the first.

# 3rd analysis

Finally, a small analysis is made, as a "discursion/diskurs" section, which groups the positions according to body type with high impact but low mobility players (Tunge), medium players with medium weight who are still semi-mobile (Medium), and the light players with high mobility but little to no impact force due to low weight (Lette). The groupings can be seen in the "Datablad". Here the thought is that if we can group the players into groups, we can include positions with less observations and have a higher, overall, amount of observations to analyse from, together with having the ability to aggregate the physical abilities of the group.

The same unsupervised and supervised analysis has been made as in the first and second analysis, the PCA analysis comes to the same conclusion where our heavy players are low in agility, with low scores in Cone, Shuttle and Forty (Higher is worse, measured in seconds) but high in BenchReps, weight and height where lighter players are more agile.

Since this dataset is grouped into 3 overlined groups, with no real definition of the actual position, this analysis can't be concluded upon and is therefore included as a "discourse/diskurs" section.

The results of the prediction algorithm is as following:

| Fit-type | K-nearest | Decision tree | Random forest |
|---|---|---|---|
| Accuracy | 86.06% | 89.9% | 88.85% |

This is, as to be expected, a very high accuracy compared to the other analysis methods.

## R conclusion

As per the overall conclusion, the R group can’t conclude that a prediction model can be made, that predicts, with a high enough accuracy, the position a player should play that is dependant on the physical characteristics of a player. However, we can conclude that the groupings of a player is highly dependant on the physical characteristics.