NFL Combine Model

Michael Snow, Han Jiang (James), Kyung Lee,
Chongyong Lyu, Abimbola Ogungbire

Project Goals

- Goals:
  - To predict where player will be drafted based on combine results
  - To predict whether or not a player will be drafted
- Usefulness
  - Would allow players to know which combine events are the most important
  - Earlier the player is drafted the greater his initial salary and generally his chance for a successful career

Data Description

- Data contains player’s combine results from the last 15 years and where the player was drafted
- The data frame contains 6218 observations and 16 variables

Forty - how long it took to run the 40 yard dash in seconds (lower the better)

Vertical - how high their vertical jump was in inches (larger the better)

BenchReps - how many reps they got on the bench press, the units are number of reps (larger the better)

BroadJump - how far their broad jump distance was in inches (larger the better)

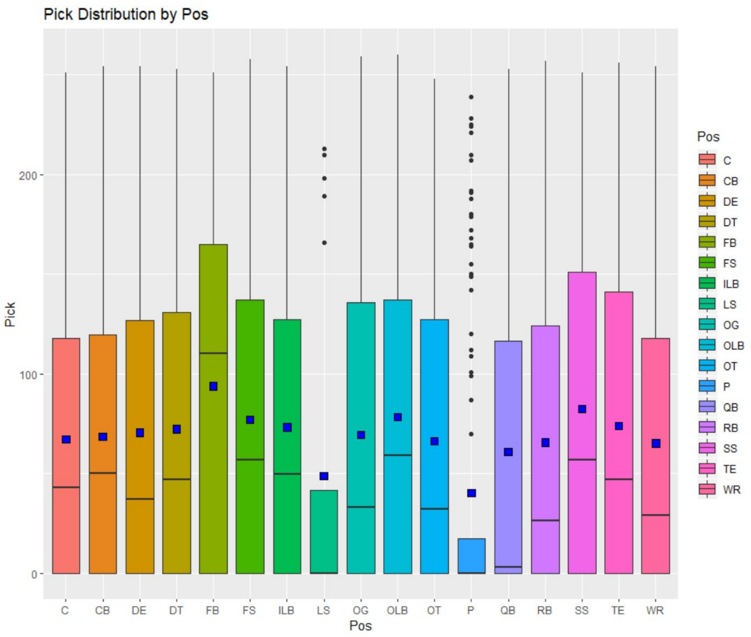
Cone - how long it took to complete the 3 cone drill in seconds (lower the better)

Shuttle - how long it took to complete the 20 yard shuttle in seconds (lower the better)

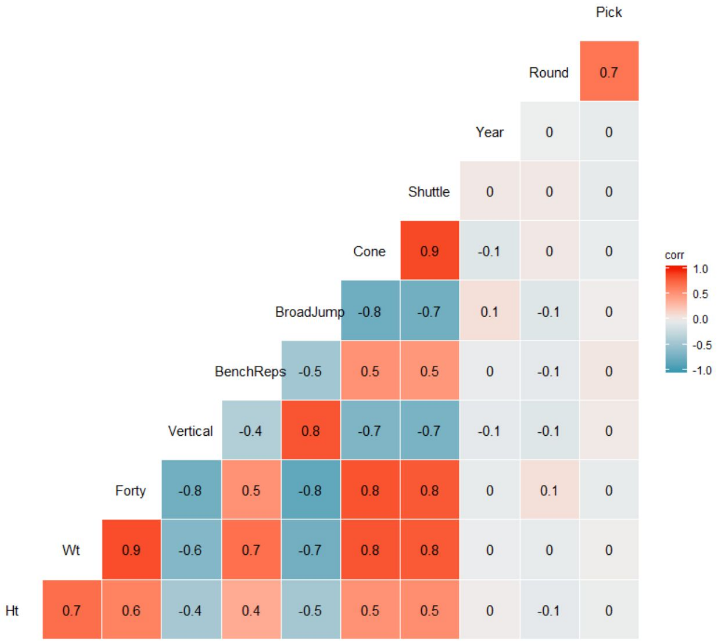
Data Preparation

- Data contained many missing values
  - Cone and Shuttle events had the most missing values
- Imputation of missing values with mean by group position.
- Listwise deletion for the impossible imputation cases (2.56% of the whole rows).

Data Exploration



* blue square inside is mean value



	Wt	Forty	Vertical	BenchReps	BroadJump	Cone	Shuttle	Year	Round	Pick
Round										0.7
Year									0	0
Shuttle								0	0	0
Cone							0.9	-0.1	0	0
BroadJump						-0.8	-0.7	0.1	-0.1	0
BenchReps					-0.5	0.5	0.5	0	-0.1	0
Vertical				-0.4	0.8	-0.7	-0.7	-0.1	-0.1	0
Forty			-0.8	0.5	-0.8	0.8	0.8	0	0.1	0
Wt		0.9	-0.6	0.7	-0.7	0.8	0.8	0	0	0
Ht	0.7	0.6	-0.4	0.4	-0.5	0.5	0.5	0	-0.1	0

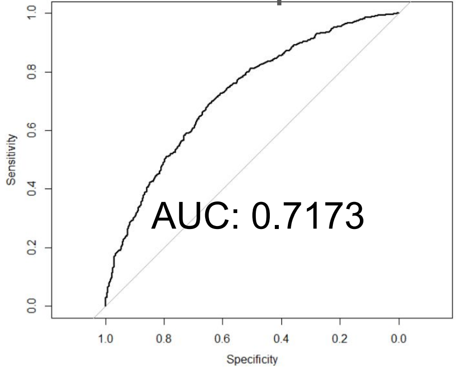
- Due to the Position P (Punter) containing mostly outliers, this position was removed from models
  - This position also does not normally participate in the combine

Linear Regression

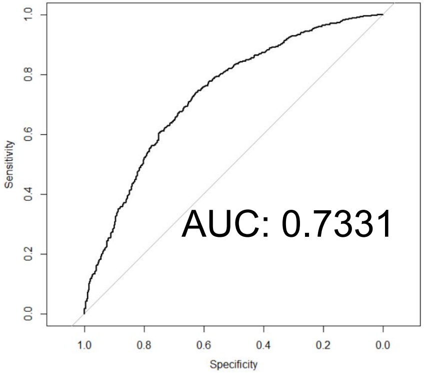
- Initial model had R2 of 0.05
    - Significant variables were Forty, Bench Reps, Broad Jump, and Shuttle
- Created separate models based on player's position
- R2 values very low
- Highest R2 values were for CB and RB positions (0.15)
    - Majority of positions had a model with an R2 of around 0.05
    - For both Forty, Broad Jump, and Cone were significant predictors
    - For RBs Vertical and Shuttle were also significant
- Backward Selection was used to choose important variables
- Interaction terms were tried but they did not improve the R2
- Due to low R2 values, we instead decided to try to predict whether a player will be drafted or not

Logistic Regression

- Response variable (Pick_Yes_No)
- Initial model
  - glm(Pick ~ Ht + Wt + Forty + Vertical +BenchReps + BroadJump +Cone +Shuttle, data=train, family = "binomial")
  - Split the data into train (0.7) and test (0.3) datasets.
- Stepwise regression Pick
  - glm(Pick ~ factor(Pos) + Wt + Forty + BenchReps + BroadJump +Cone +Shuttle, data=train, family = "binomial")
- Initial model Vs Stepwise Pick



AUC: 0.7173

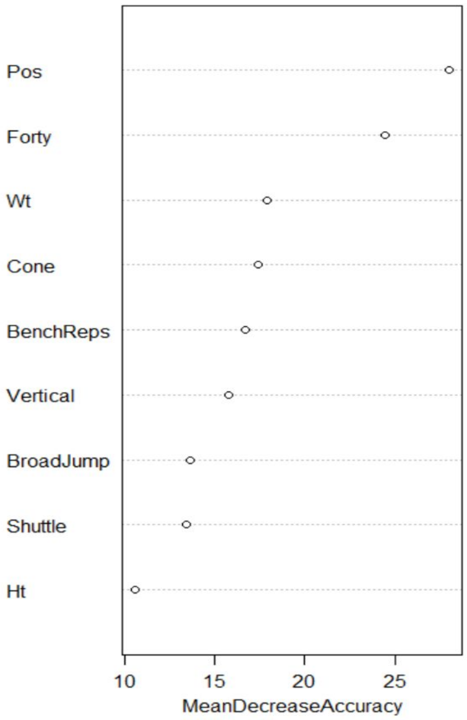


AUC: 0.7331

Decision Tree

- Response variable (Pick_Yes_No)
- Initial model with 9 explanatory variables
    - tree_ROSE <-rpart(Pick_Yes_No ~ Pos + Ht + Wt + Forty + Vertical + BenchReps + BroadJump + Cone + Shuttle, data = train_ROSE, method = "class")
- Variable selection approach
    - What variable to select or drop on based MeanDecreaseAccuracy of Random Forest
- Modeling step
    - Split train-test (with 70% split ratio)
    - Oversampling
    - Build models by excluding likely less important variables (3 variants using oversampling, adjustment of CP, and pruning for each model)
    - Compare AUCs
    - Visualize ROC curves

Decision Tree (continued)

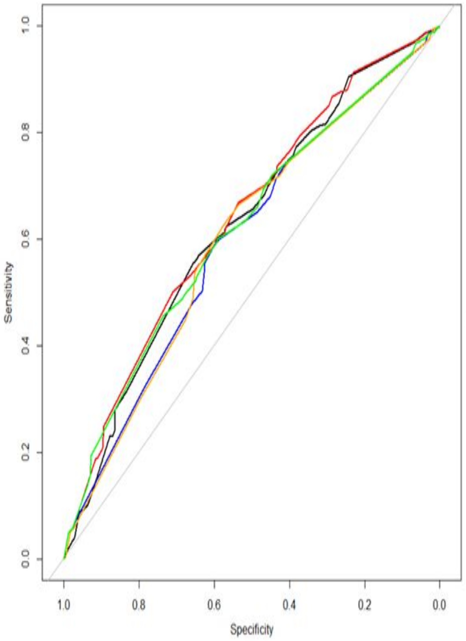


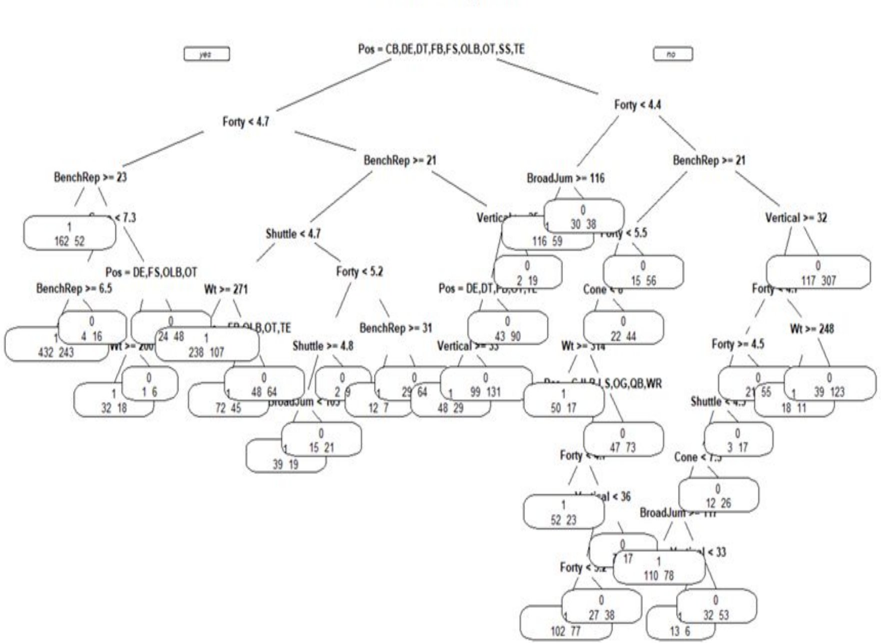
Drop variables
one by one
Based on Mean Decrease in Accuracy

No	Model	Over-Samplin g (AUC)	Adjustment CP (AUC)	Pruning (AUC)
1	(9 variables)	0.5932	0.6273	0.6273
2	(8 variables) Drop Ht	0.5932	**0.6385**	0.6386
3	(7 variables) Drop Shuttle	0.597	0.6195	0.6195
4	(6 variables) Drop BroadJump	0.597	0.5984	0.5984
5	(5 variables) Drop Vertical	0.597	0.602	0.602
...	...	...	...	...

- Pruning does not make any improvement

Decision Tree (continued)




Model 2 - CP adjusted

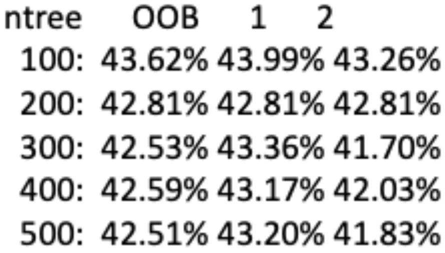
Selected model

- Model #2 (red)
- AUC : **0.6385**
- Eight explanatory variables matter (Pos, Wt, Forty, Vertical, BenchReps, BroadJump, Cone and Shuttle)

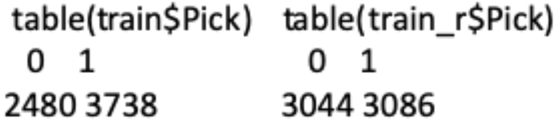
Random Forest Classification

Pick Model OBB

OOB estimate of error rate: 42.51%



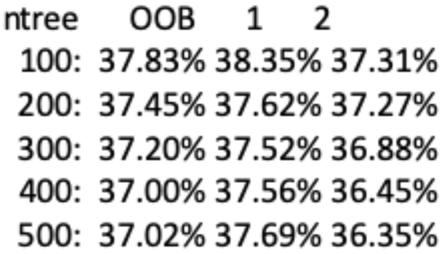
```
ntree    OOB     1      2
  100:  43.62% 43.99% 43.26%
  200:  42.81% 42.81% 42.81%
  300:  42.53% 43.36% 41.70%
  400:  42.59% 43.17% 42.03%
  500:  42.51% 43.20% 41.83%
```



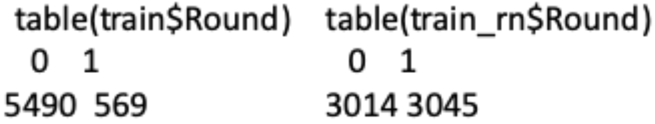
```
 table(train$Pick)   table(train_r$Pick)
   0    1                 0    1
 2480 3738              3044 3086
```

Round Model OBB

OOB estimate of error rate: 37.02%

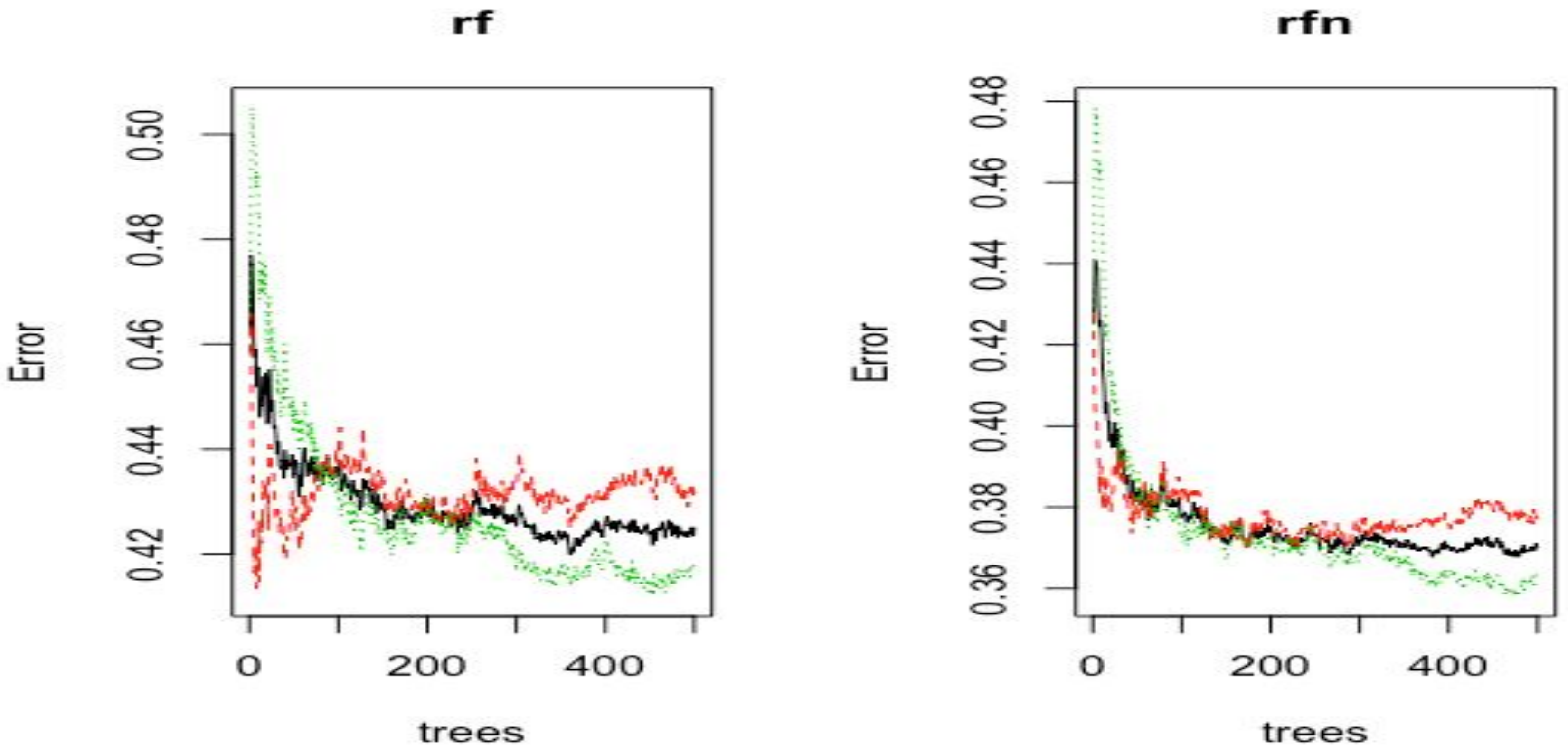


```
ntree    OOB     1      2
  100:  37.83% 38.35% 37.31%
  200:  37.45% 37.62% 37.27%
  300:  37.20% 37.52% 36.88%
  400:  37.00% 37.56% 36.45%
  500:  37.02% 37.69% 36.35%
```



```
 table(train$Round)   table(train_rn$Round)
   0    1                 0    1
5490  569              3014 3045
```

Optimal numbers of trees (left:Pick, right:Round)

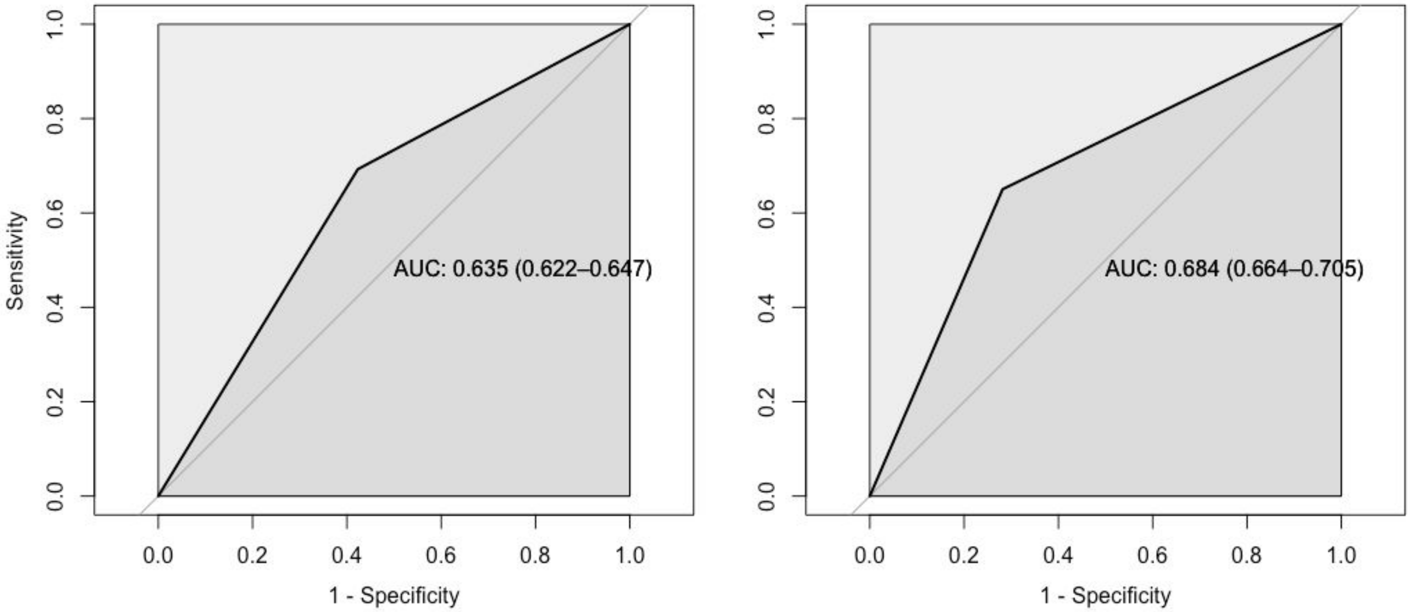


Comments: After increasing tree units above 300 units, the error of the model seemed to be varied and reduced (Green line refers to the lowest OBB error rate).

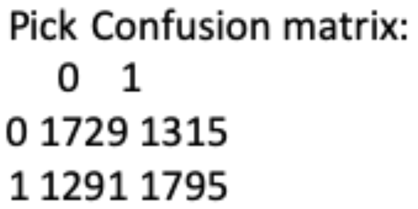
Random Forest Model Result

Left: Pick Prediction

Right: Round Prediction

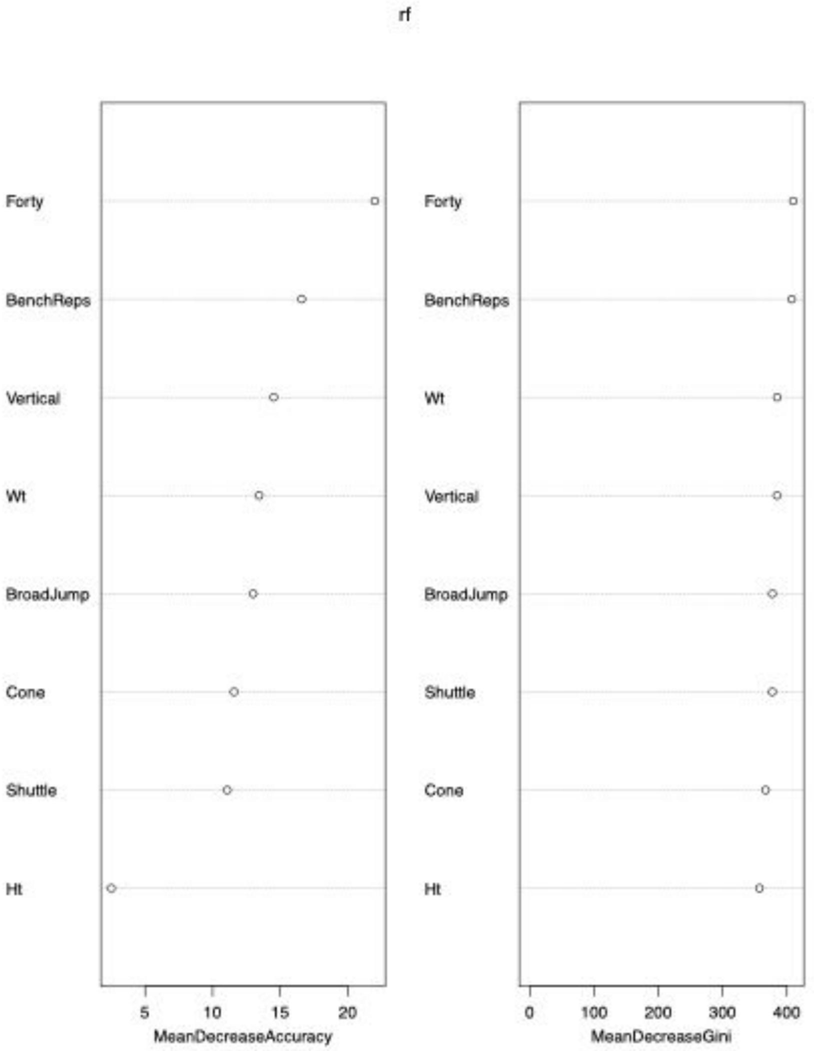


Pick Confusion Matrix

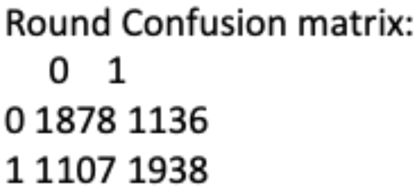


```
Pick Confusion matrix:
    0    1
0 1729 1315
1 1291 1795
```

- Sensitivity: 0.5725
- Specificity:0.5772
- Accuracy:0.5749

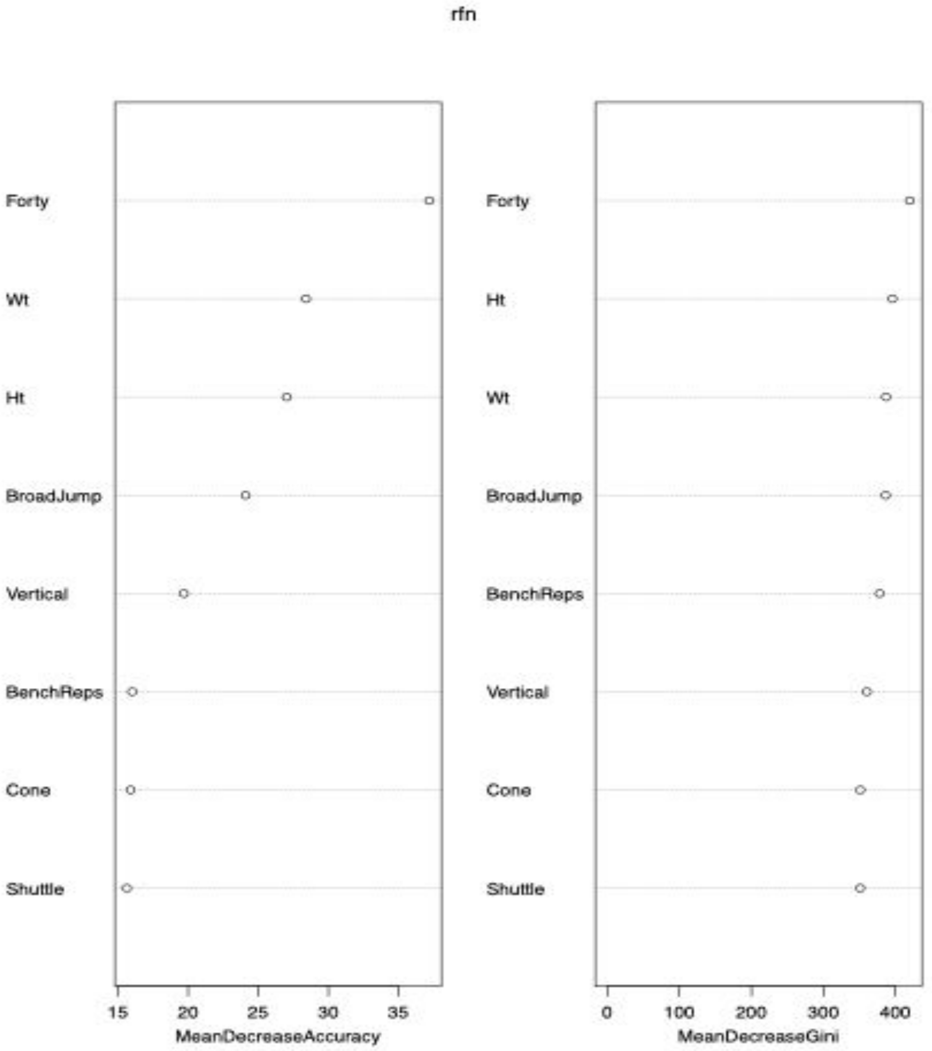


Round Confusion Matrix



```
Round Confusion matrix:
   0    1
0 1878 1136
1 1107 1938
```

- Sensitivity: 0.6291
- Specificity:0.6304
- Accuracy:0.6298



Summary of results

- Based on our analysis, the most important predictor of draft pick is forty taking into consideration of the player positions overall.
    - Logistic Regression is the best model
- Linear Regression model for overall data was 0.05
    - When seperated by position, the positions RB and CB had their R2 increase to 0.15
        - The other positions' R2 remained the same
- Logistic Regression model has the best performance with AUC of 0.73
- The best performance of Decision Tree model comes with AUC of 0.64
    - Adjustment cp technique contributes model performance
- Random Forest model obtains the most predicting power when number of trees are set beyond 400 units (Pick_AUC: 0.635; Round_AUC:0.684; Top_1 important predictor: “Forty”)

Conclusion

Models are not good for predicting Pick but are decent for predicting whether a player will be drafted or not.

Future Improvement

- Due to the high amount of unexplained variability in predicting Pick, it is pertinent that we address the limiting factors that may reduce it. Injury, character and motivation can affect the performance of a player. Factors as such are not considered. To make better predictions, further study needs to consider taking in new attributes to re-investigate multicollinearity.