# NFL: Predict Running Back Yardage

*Cedric Williams (705620756) & Kevin Jackson (730288402)*

*12/9/2019*

## PROBLEM

Every college and professional football team is looking for a competitive advantage, some insight to help their team perform better. The problem is the game of football involves many facets and it is hard for humans to extract insights with a game that has so many moving parts. In football, there are 22 men on the field that may or may not contribute to your victory. Using Machine Learning and tracking sensors, teams potentially can model the opponent's defense and gain valuable insight. Any insight gained from machine learning is valuable to many organizations. We will create a simple example that answers a key question, based on the number of defenders in the box, can we predict the number of yards gained by a running back?

## APPLICATION

Predicting the running backs output based on the number of defensive players at the line of scrimmage, would be helpful and assist coach's with in-game planning or designing plays. With the proper model, teams could predict their running backs performance and alter their game strategy. By modeling running backs' yard output with features such as defenders in the box and other characteristics, coaches could exploit scenarios by going "no-huddle" and keeping the same number of defenders at the line. This is just one example of how the offense would utilize this algorithm. If the defense had this model, they would be able to tailor yard stopping plays on fourth and inches.

## MOTIVATION

Big Data and Analytics are an integral part of the NFL experience and way beyond marketing. Kaggle.com and the NFL teamed up again for the Second Annual Big Data Bowl. This year's main question is predicting how many yards the running back would gain on a single play. We will need an ensemble algorithm predicting different components to win this competition. Our objective focuses on the number of defenders at the line dictating the running back's performance.

## SURVEY OF RELATED WORK

Sports Analytics is a hot topic and has been for a while in basketball, baseball or even soccer. But the NFL recently tagged players' uniforms and game balls so there isn't a huge amount of related work at this time. This type of data has only been available for the past two years.

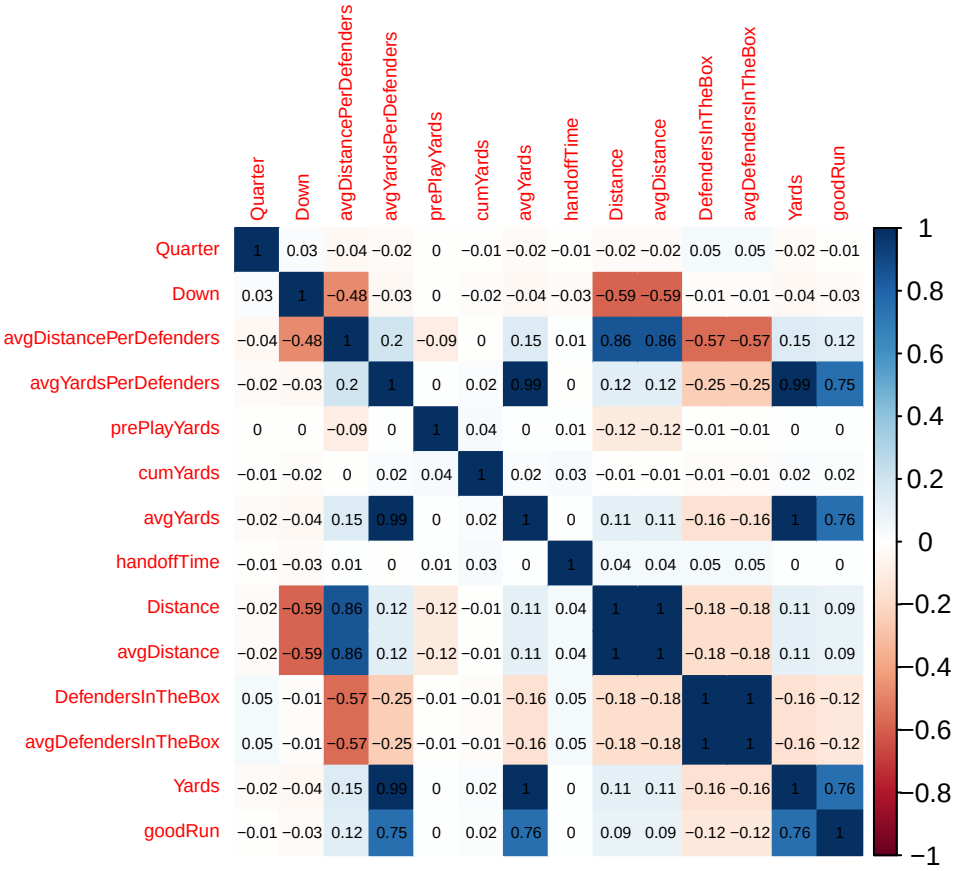
## APPROACH TO THE MODEL

We first examined the data set and eliminated columns/rows that weren't useful for our analysis. Since we are looking at running back output, we only need the running back row data. Each row has meta-data about the play so eliminating the other rows did not change the data. In the end, our aggregated list of columns was: Nfld, DisplayName, PlayerHeight, PlayerWeight, PlayerBirthDate, Position, GameId, PlayId, Turf, Quarter, Down, FieldPosition, YardLine, Distance, OffenseFormation, OffensePersonnel, DefendersInTheBox, DefensePersonnel, PlayDirection, TimeHandOff, TimeSnap, Yards, LastRunPlayYards, and CumYards.

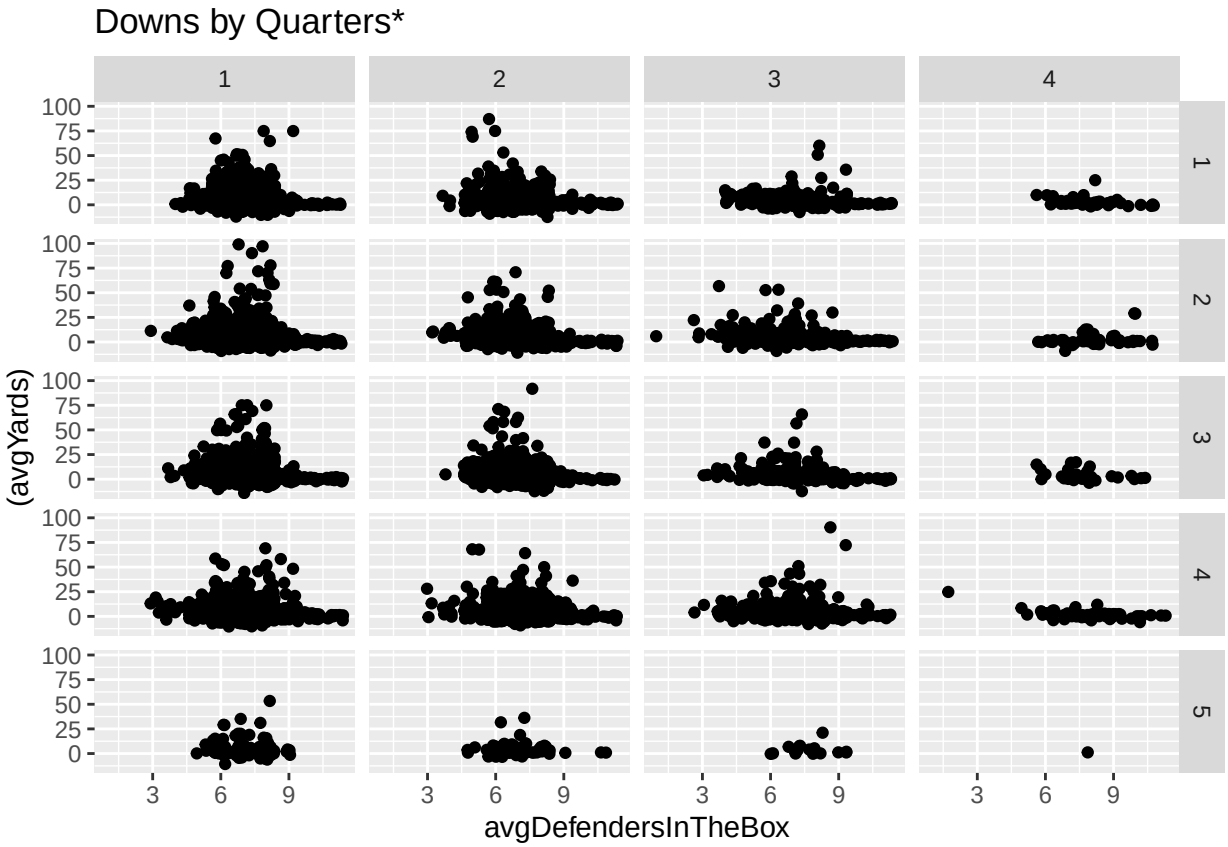
## EXPLORATORY ANALYSIS

We conduct some feature engineering and create additional features such as avgYardsPerDefender, avgDistancePerDefenders, avgDefendersInTheBox, and lastly a measure called goodRun for any rush that earned a first down or went over 5 yards. We explored the dimensions and metrics to discover any correlations. There was some redundant data that will will eliminate in the modeling stage.

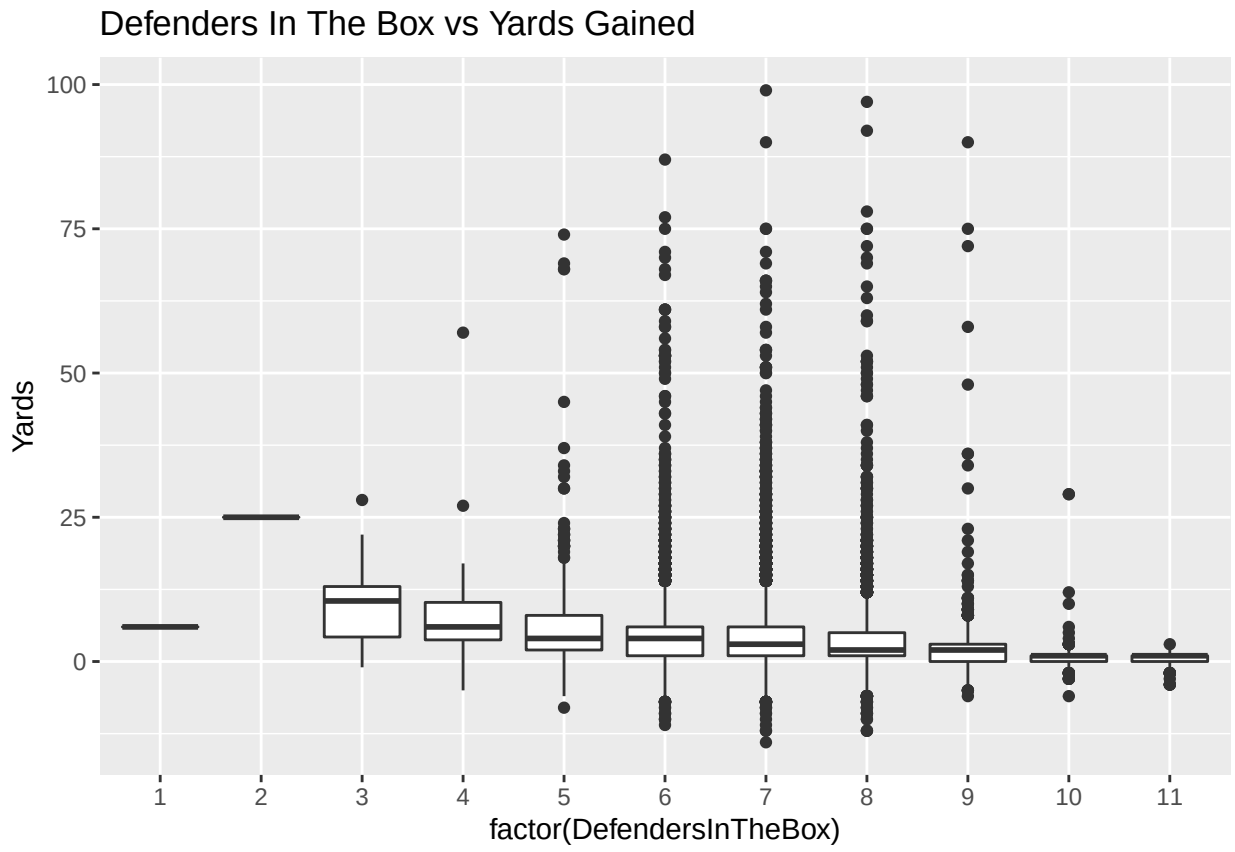
Below is our correlation matrix. Our assumption is the number of defenders in the box determines running back success. There was limited correlation between good runs and number defenders in the box. Metrics had higher correlation when they compared distance with downs or defenders.




We also compared defenders in the box against yards gained but faceted by the down and the quarter. The 5th quarter is overtime. The data clustered and seemed to follow a curve but wasn't very linear for the earlier downs. There was more linear formation as the downs increased to 4th down. It was evident that the distance and downs had a major influence.




Each time an additional defender was added to the box, the median yardage gained was decreased. For our analysis, we're only looking at running plays. Increasing the defenders in the box exposes the defense against the big pass play.




**PREDICTIVE MODEL**

Our model was a linear regression. We started with more features but realized that most features were not improving the model.

```
##
## Call:
## lm(formula = Yards ~ DefendersInTheBox + Distance + cumYards +
##     Down + Quarter + prePlayYards + handoffTime, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.365  -3.191  -1.206   1.295  94.171
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.227e+00  5.193e-01  13.916  < 2e-16 ***
## DefendersInTheBox -6.317e-01  5.348e-02 -11.811  < 2e-16 ***
## Distance           9.834e-02  1.659e-02   5.929 3.11e-09 ***
## cumYards           3.810e-04  9.735e-05   3.913 9.14e-05 ***
## Down               2.583e-02  8.831e-02   0.293   0.7699
## Quarter           -1.411e-02  4.403e-02  -0.321   0.7486
## prePlayYards       1.206e-02  7.755e-03   1.555   0.1199
## handoffTime        2.369e-01  1.072e-01   2.211   0.0271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.513 on 16724 degrees of freedom
## Multiple R-squared:  0.01628,    Adjusted R-squared:  0.01586
## F-statistic: 39.53 on 7 and 16724 DF,  p-value: < 2.2e-16
```

For our first model, we used more parameters than suggested by the high Pr() score, and left within the model Down, Quarter, and prePlayYards. Our accuracy was 0.391988 and based on limited data, we're pleased. With limited data and time to create more features, we're unable to increase the accuracy at this point.

We removed the high probability features and re-ran the model. There was a minimal improvement.

```
##
## Call:
## lm(formula = Yards ~ DefendersInTheBox + Distance + cumYards +
##     handoffTime, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.232  -3.191  -1.202   1.285  94.188
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.335e+00  4.376e-01  16.762  < 2e-16 ***
## DefendersInTheBox -6.359e-01  5.287e-02 -12.026  < 2e-16 ***
## Distance           9.516e-02  1.407e-02   6.765 1.38e-11 ***
## cumYards           3.860e-04  9.721e-05   3.971 7.19e-05 ***
## handoffTime        2.371e-01  1.071e-01   2.213   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.513 on 16727 degrees of freedom
## Multiple R-squared:  0.01613,    Adjusted R-squared:  0.01589
## F-statistic: 68.54 on 4 and 16727 DF,  p-value: < 2.2e-16
```

When we reduced our model to the most significant parameters, we see a slight accuracy improvement of 0.3922878 over 0.391988 but not enough to move the needle. Overall, we need more features to better describe and predict running back yardage. Defenders in the box are the most prominent feature but by itself, can't predict the running back's outcome.

## CONCLUSION

In conclusion, predicting rushing yards is extremely complicated and even with a simple linear regression model, we were only able to nearly reach 40% accuracy. We intend to add more features and have a goal to surpass 60% accuracy.

## REFERENCES

Kaggle & NFL Big Data Bowl Challenge: https://www.kaggle.com/c/nfl-big-data-bowl-2020

Cedric Williams Github: https://github.com/ceo4ced/nflRunningBacks