NFLProject

CW & KJ

12/9/2019

##Problem Each professional football team is looking for an advantage and so is every gambler out there as well. Each of the 32 NFL teams is looking for an edge and potentially solving that edge just might bring a Lombardi trophy across this state yonder. Any insight gained from machine learning is valuable to many organizations. We will create a toy example that ponders, based on how many defenders are in the box, can we predict the number of yards gained by a running back?

##Application Improving either the offense's yard gained or helping the defense restrict yards gained gets at the core of a team's run offense or defense. By modeling running backs yard output by features such as defenders in the box AND ########, 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.

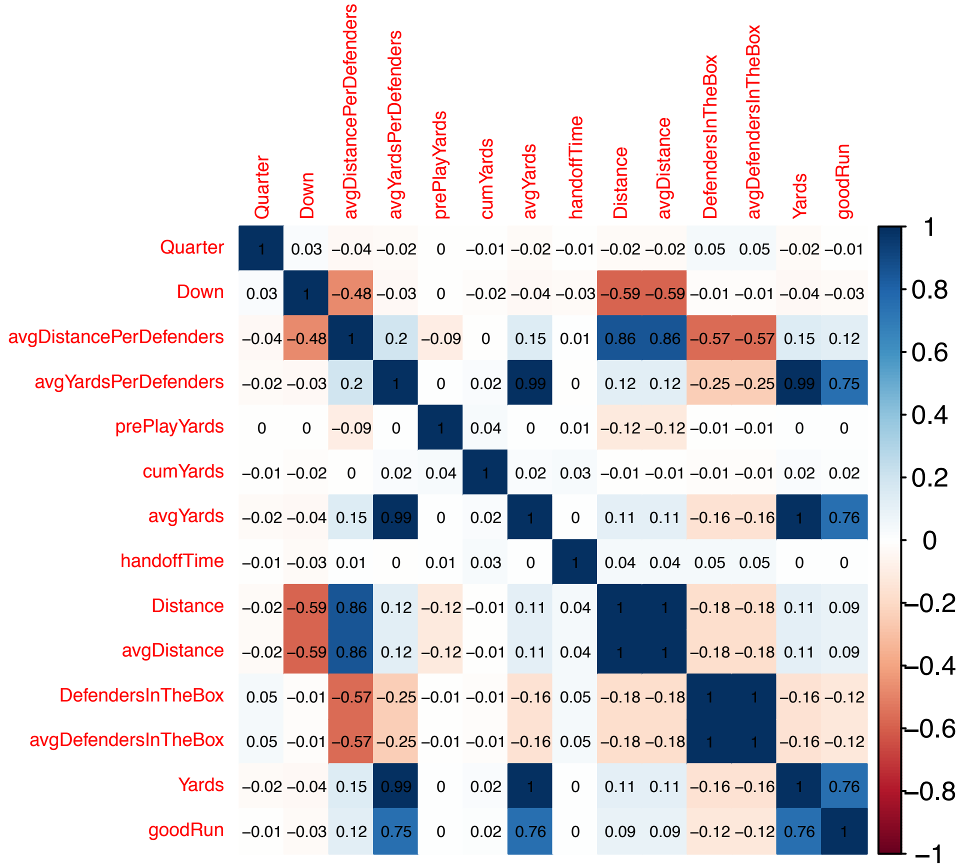
##Motivation Big Data and ANalytics are an integral part of the NFL experience and beyond marketing. Kaggle.com and the NFL teamed up again for the Second Annual Big Data Bowl. This year's main question is predicting how many yards the running back would gain on a single play. To win this competition, we'll need an ensemble algorithm predicting different components. But for this project, we will focus on the defenders and running back yards gained.

##Survey of Related Work Sports Analytics is a hot topic and has been for a while for 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 model We first took the data set and eliminated which columns/rows 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 metadata 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.

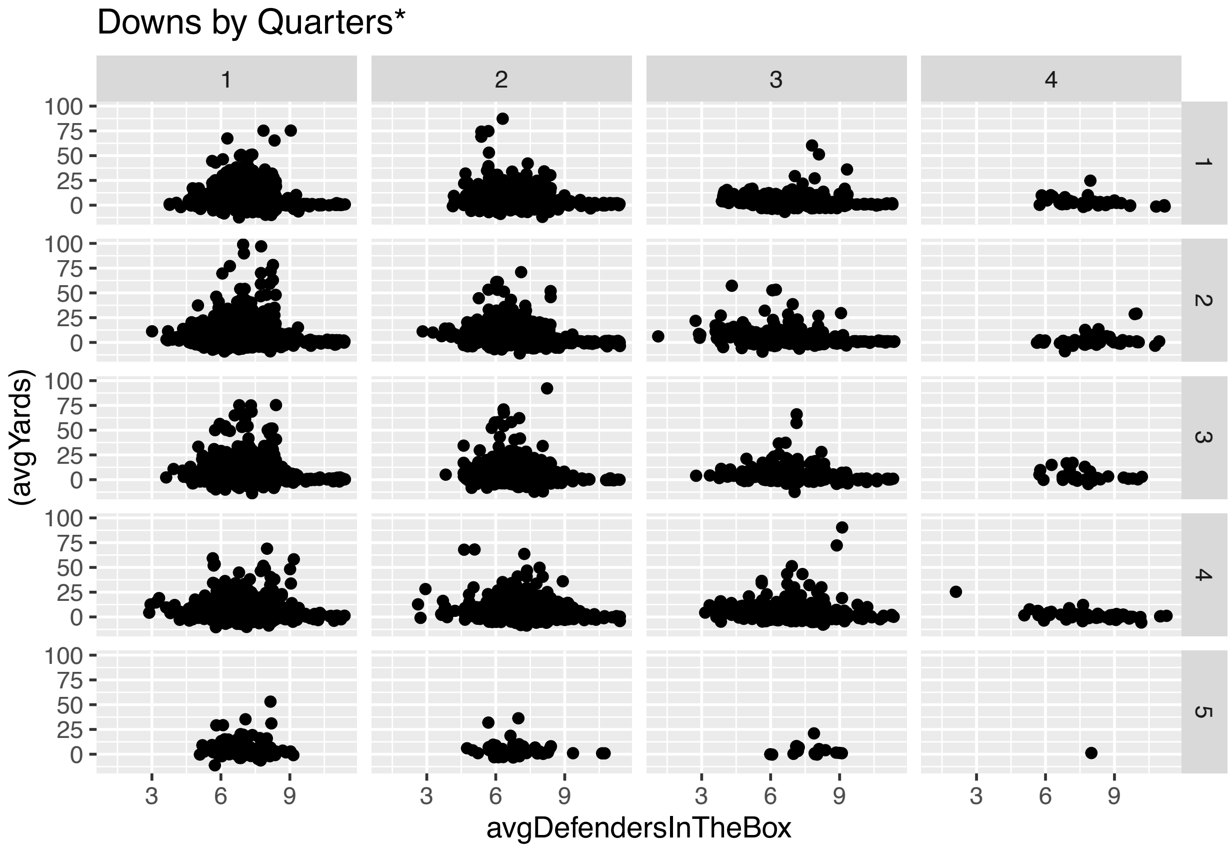
#Exploratory Analysis We conduct some feature engineering and create additional features such as avgYardsPerDefender, avgDistancePerDefenders, avgDefendersInTheBox, and lastly a measure called goodRun for any rush that earned a first down or went over 5 yards.

Below is our correlation matrix.



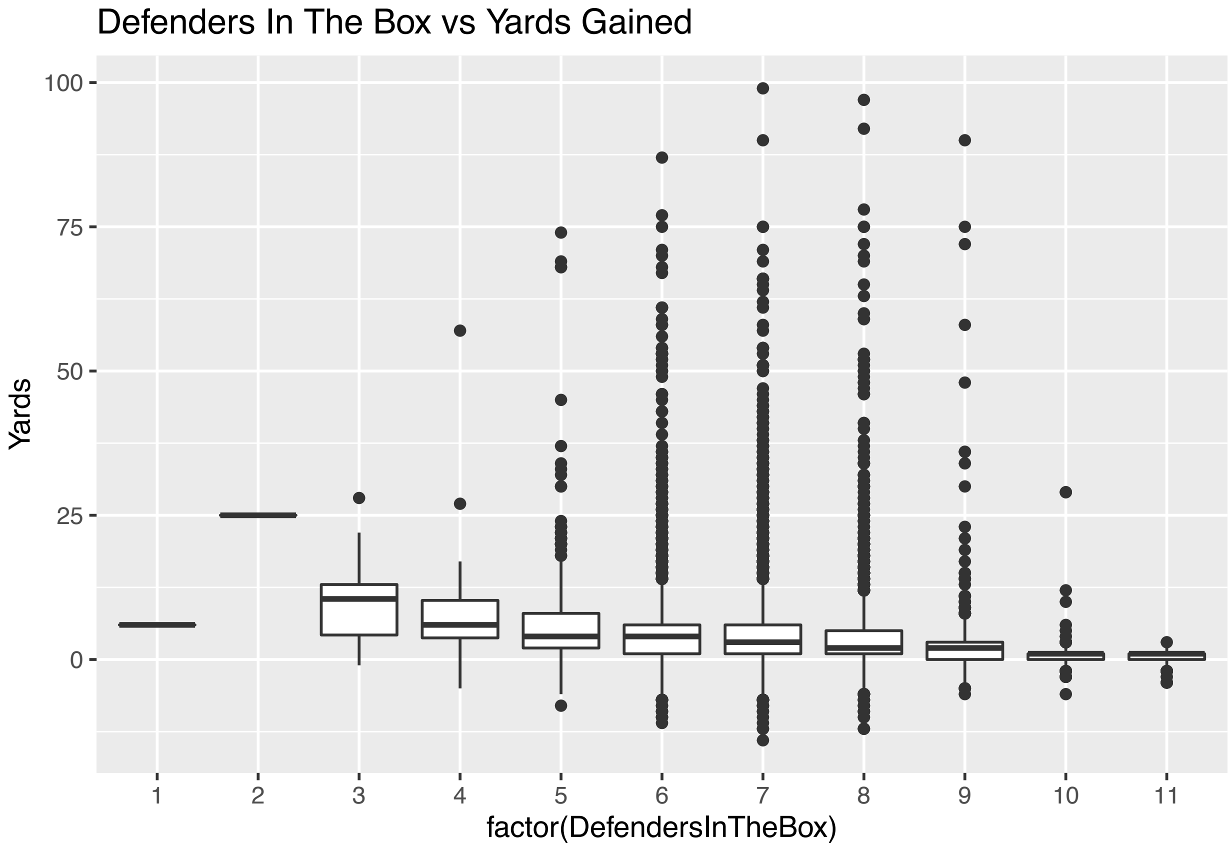
	Quarter	Down	avgDistancePerDefenders	avgYardsPerDefenders	prePlayYards	cumYards	avgYards	handoffTime	Distance	avgDistance	DefendersInTheBox	avgDefendersInTheBox	Yards	goodRun
Quarter	1	0.03	−0.04	−0.02	0	−0.01	−0.02	−0.01	−0.02	−0.02	0.05	0.05	−0.02	−0.01
Down	0.03	1	−0.48	−0.03	0	−0.02	−0.04	−0.03	−0.59	−0.59	−0.01	−0.01	−0.04	−0.03
avgDistancePerDefenders	−0.04	−0.48	1	0.2	−0.09	0	0.15	0.01	0.86	0.86	−0.57	−0.57	0.15	0.12
avgYardsPerDefenders	−0.02	−0.03	0.2	1	0	0.02	0.99	0	0.12	0.12	−0.25	−0.25	0.99	0.75
prePlayYards	0	0	−0.09	0	1	0.04	0	0.01	−0.12	−0.12	−0.01	−0.01	0	0
cumYards	−0.01	−0.02	0	0.02	0.04	1	0.02	0.03	−0.01	−0.01	−0.01	−0.01	0.02	0.02
avgYards	−0.02	−0.04	0.15	0.99	0	0.02	1	0	0.11	0.11	−0.16	−0.16	1	0.76
handoffTime	−0.01	−0.03	0.01	0	0.01	0.03	0	1	0.04	0.04	0.05	0.05	0	0
Distance	−0.02	−0.59	0.86	0.12	−0.12	−0.01	0.11	0.04	1	1	−0.18	−0.18	0.11	0.09
avgDistance	−0.02	−0.59	0.86	0.12	−0.12	−0.01	0.11	0.04	1	1	−0.18	−0.18	0.11	0.09
DefendersInTheBox	0.05	−0.01	−0.57	−0.25	−0.01	−0.01	−0.16	0.05	−0.18	−0.18	1	1	−0.16	−0.12
avgDefendersInTheBox	0.05	−0.01	−0.57	−0.25	−0.01	−0.01	−0.16	0.05	−0.18	−0.18	1	1	−0.16	−0.12
Yards	−0.02	−0.04	0.15	0.99	0	0.02	1	0	0.11	0.11	−0.16	−0.16	1	0.76
goodRun	−0.01	−0.03	0.12	0.75	0	0.02	0.76	0	0.09	0.09	−0.12	−0.12	0.76	1

```
plots()
g + geom_jitter(aes(y=(avgYards),x=avgDefendersInTheBox))+facet_grid(Quarter~Down)+ggtitle("Downs by Qua
```



Including Plots

You can also embed plots, for example:



```
##
## Call:
## lm(formula = Yards ~ DefendersInTheBox + Distance + cumYards +
##     Down + Quarter + prePlayYards + handoffTime, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.213  -3.205  -1.228   1.298  94.239
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.106e+00  5.178e-01  13.724  < 2e-16 ***
## DefendersInTheBox -6.124e-01  5.324e-02 -11.502  < 2e-16 ***
## Distance           9.094e-02  1.619e-02   5.616 1.98e-08 ***
## cumYards           3.512e-04  9.852e-05   3.564 0.000366 ***
## Down               1.043e-01  8.754e-02   1.192 0.233274
## Quarter           -2.160e-02  4.398e-02  -0.491 0.623363
## prePlayYards       8.923e-03  7.728e-03   1.155 0.248284
## handoffTime        2.300e-01  1.072e-01   2.146 0.031859 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.503 on 16724 degrees of freedom
## Multiple R-squared:  0.01469,    Adjusted R-squared:  0.01428
## F-statistic: 35.63 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 in Down, Quarter, and prePlayYards. Our accuracy was 0.3960314 and based on limited data, we're pleased. With limited data and time to create more features, we're unable to increaes the accuracy at this point.

```
##
## Call:
## lm(formula = Yards ~ DefendersInTheBox + Distance + cumYards +
##     handoffTime, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.261  -3.203  -1.222   1.275  94.230
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.413e+00  4.357e-01  17.013  < 2e-16 ***
## DefendersInTheBox -6.235e-01  5.261e-02 -11.851  < 2e-16 ***
## Distance           8.037e-02  1.382e-02   5.817 6.09e-09 ***
## cumYards           3.515e-04  9.843e-05   3.571 0.000356 ***
## handoffTime        2.303e-01  1.072e-01   2.149 0.031646 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.503 on 16727 degrees of freedom
## Multiple R-squared:  0.01453,    Adjusted R-squared:  0.0143
## F-statistic: 61.66 on 4 and 16727 DF,  p-value: < 2.2e-16
```

When we reduced our model to the most significant parameters and make it simple, we see a slight accuracy improvement of 0.3962364 over 0.3960314 but not enough to move the needle.

In conclusion, predicting rushing yards is extremely complicated and even with a simple model, we were only able to almost reach 40% accuracy.