WVU Football - 2019 vTexas

Jeremy Harris

April 28, 2020

WVU Football Analysis

I'm creating this analysis as a pipeline to view WVU Football information. I have started with just a single game here and the data is limited. I plan to expand this into a Shiny App so a user can select the year and game that they are interested in and have the results automatically appear.

In addition, I plan to compile multiple years for a head coach and try to gain further insight into tendencies related to the game current game situation. This is just a starting point and will be built upon further.

Texas @ WVU - 2019

The game I'm using to setup the model is week 6 of the 2019 season when Texas visited WVU. Once I get the analysis fully setup, this should work for any team that WVU played.

The Data

I'm using this website: https://collegefootballdata.com to gather data for all plays that took place during the game. They have an API located here: https://api.collegefootballdata.com/api/docs/?url=/api-docs.json#/plays/getPlays that I am using to pull the data directly into my file.

Clean Up Data

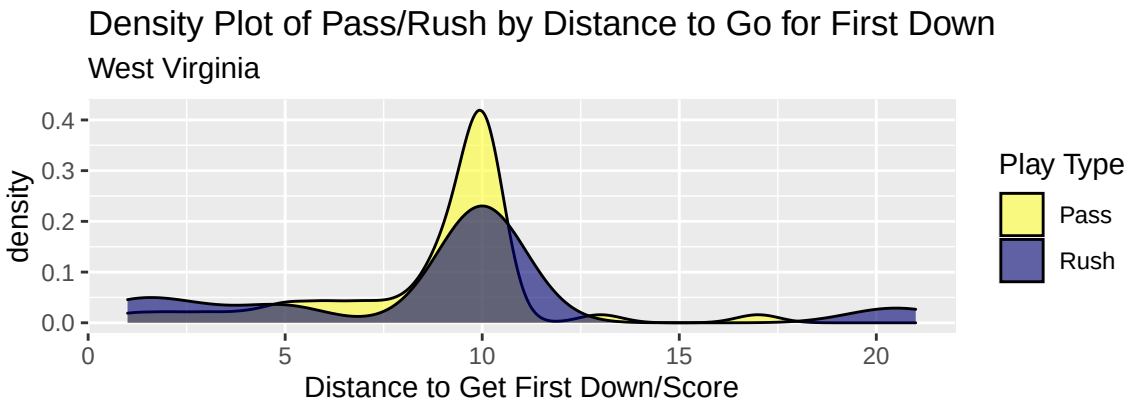
I take the data and drop fields that aren't needed, and rearrange the data into an order that makes sense - then I sort by period then minutes then seconds. Most of my code I dropped from the pdf printout but it available on my github page (http://github.com/jeremy-harris). I left this here to show some of what is taking place behind the scenes.

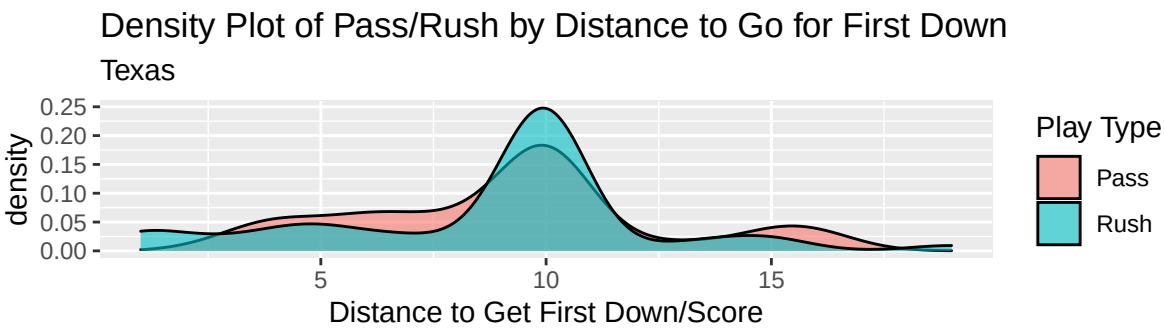
```
#convert score to be associated with team, not offense/defense
data_in <- data_in %>%
  mutate(score_WVU = ifelse(offense == "West Virginia", offense_score, defense_score),
         opp_score = ifelse(offense == opp, offense_score, defense_score)) %>%
  mutate(clock.minutes = clock[[1]], clock.seconds = clock[[2]])

#group by period, then minute, then second
data_in <- data_in %>%
  select(-drive_id, -offense_score, -defense_score, -yards_to_goal, -offense_conference, -id,
         -clock, -defense_conference, -defense, -ppa) %>%
  select(score_WVU, opp_score, period, clock.minutes,
         clock.seconds, offense, yard_line, down, distance, yards_gained, play_type, play_text) %>%
  arrange(period, -clock.minutes)
```

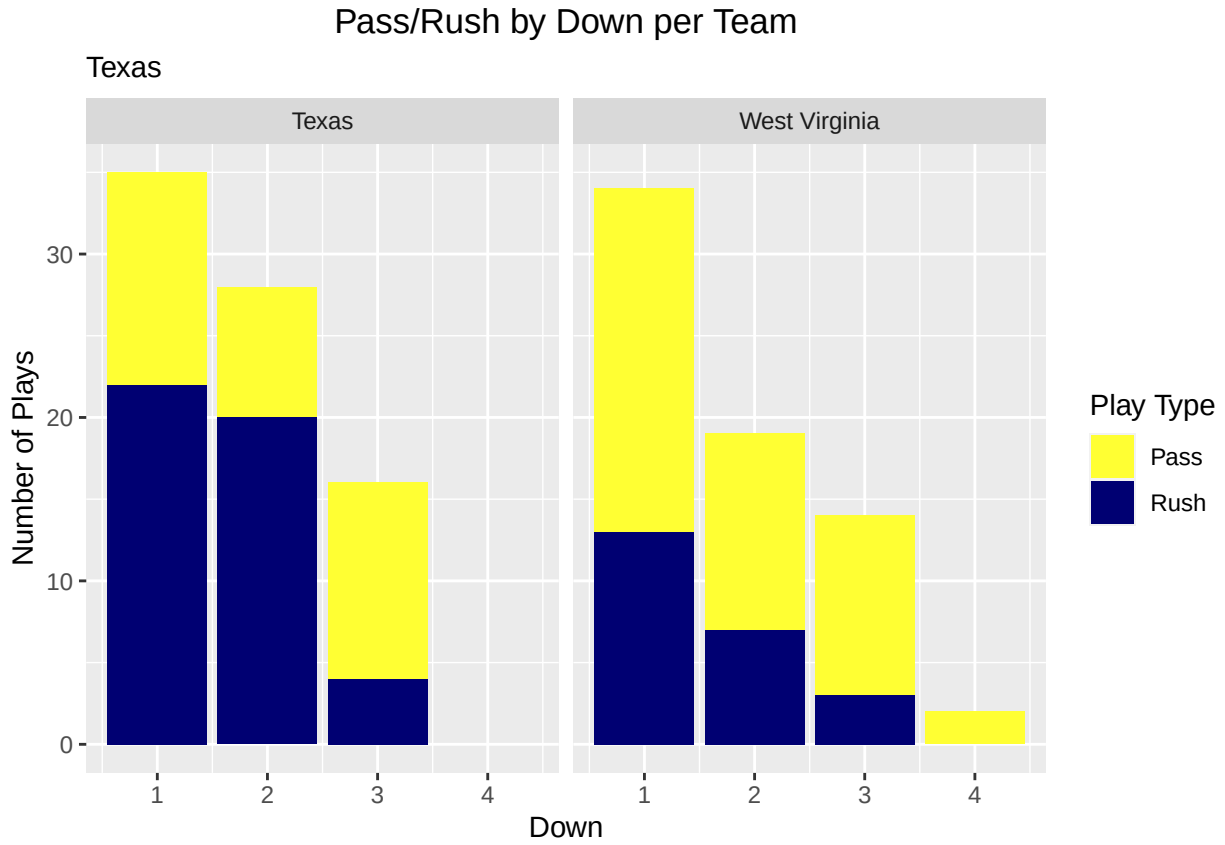
Visualize The Data

I create a few different plots here to show things that might be interesting. I don't have tons of granular detail like formations, blitzing, injuries, etc. - but I can show basic game stats which is what I try to do here.





This next plot shows the type of play by down per team.



So, clearly Texas favored the run and WVU favored the pass. Also, since punting and field goals have been removed from this graph - we can see that Texas didn't go for it on 4th down during this game.

Statistical Analysis

I want to take a look at the probability of a given play type using all variables. As this is a single game, there isn't much data to use in this instance. However, a longer term project can compile data for an entire season (or coaches career, etc.) to show tenancies given the data in the dataset. Again, this can be combined with things like formation type of offense and defense to get an even better insight.

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

I've setup models (Random Forest & Class Trees) to predict both WVU and the opponent's play to be a Run or Pass. Let's compare our models and see how they perform.

```
##
## Random Forest - WVU
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Pass Rush
##       Pass    4    2
##       Rush    0    0
##
##                Accuracy : 0.6667
##                  95% CI : (0.2228, 0.9567)
##     No Information Rate : 0.6667
##     P-Value [Acc > NIR] : 0.6804
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.4795
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.6667
##          Neg Pred Value :    NaN
##              Prevalence : 0.6667
##          Detection Rate : 0.6667
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Pass
##


##
## Random Forest - opp

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Pass Rush
##       Pass    1    0
##       Rush    2    4
##
##                Accuracy : 0.7143
##                  95% CI : (0.2904, 0.9633)
##     No Information Rate : 0.5714
##     P-Value [Acc > NIR] : 0.3593
##
##                   Kappa : 0.3636
##
##  Mcnemar's Test P-Value : 0.4795
##
##             Sensitivity : 0.3333
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.6667
##              Prevalence : 0.4286
##          Detection Rate : 0.1429
##    Detection Prevalence : 0.1429
```

```
##       Balanced Accuracy : 0.6667
##
##        'Positive' Class : Pass
##
```

What The Stats Mean

From my limited data, I was able to show a 66% accuracy predicting the play type when WVU had the ball and a 71% accuracy predicting the play type when the opponent had the ball. However, that's a little ambitious and pretty generic given that my test set of data only had 6 or 7 plays per team.

For WVU I was able to predict 4 pass plays correctly but missed 2 rushing plays. In effect, my model predicted all plays to be passing which is a decent approach considering how much WVU passed in this game vs. rushing - but I wouldn't consider that necessarily a "prediction.".

For Texas I was able to see a little more accuracy and my model correctly predicted the pass play as well as 4 rushing plays. This time the model favors rushing here because Texas ran the ball more than passing. This model predicted 2 rushing plays that were actually pass plays.

What's Next?

I think that looking at more games with more plays will provide better results, or at least results that even out. I think I'd like to show more plots with more analysis automatically.

Two things I really want to do is give the user access to change the team and game and automatically show the same results. The next thing I want to do is some sort of animation - maybe show the number scores animate over time... or, for a really tricky problem (that may not really be anything other than "cool") - animate the position of the ball for the entire game using the yard-line and yards gained per play.

Let Me Know!

If you have a stat that you want to see or something you'd like to know... please email me. I'll do my best to build it in the model.

Thanks!

jeremy.scott.harris@gmail.com