R Notebook

Alex Baker, James Kitch, Jackson Smith, Matthew Sheridan

Introduction:

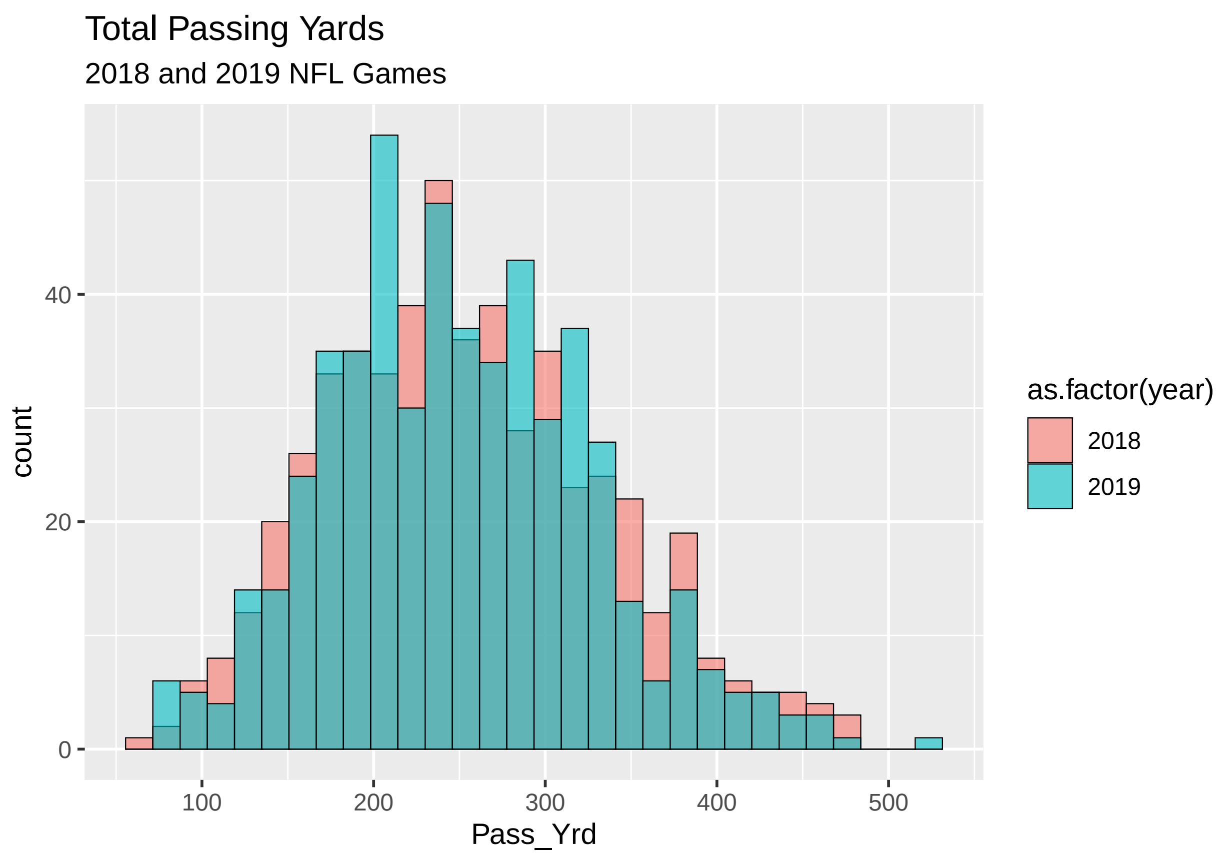
Our data comes from the nflfastR package, which contains accumulated play-by-play data from 1999 through 2021, with additional predictors beginning in 2006. We hope to create an efficient, streamlined model capable of predicting the score differential of an NFL game. Furthermore, we also want to be able to use what we gain from our models to help make insightful analyses and predictions in other scenarios. Some predictors will impact only one of these models while others will prove to be far more significant than widely recognized: it all depends on which refining method produces the most successful and significant model.

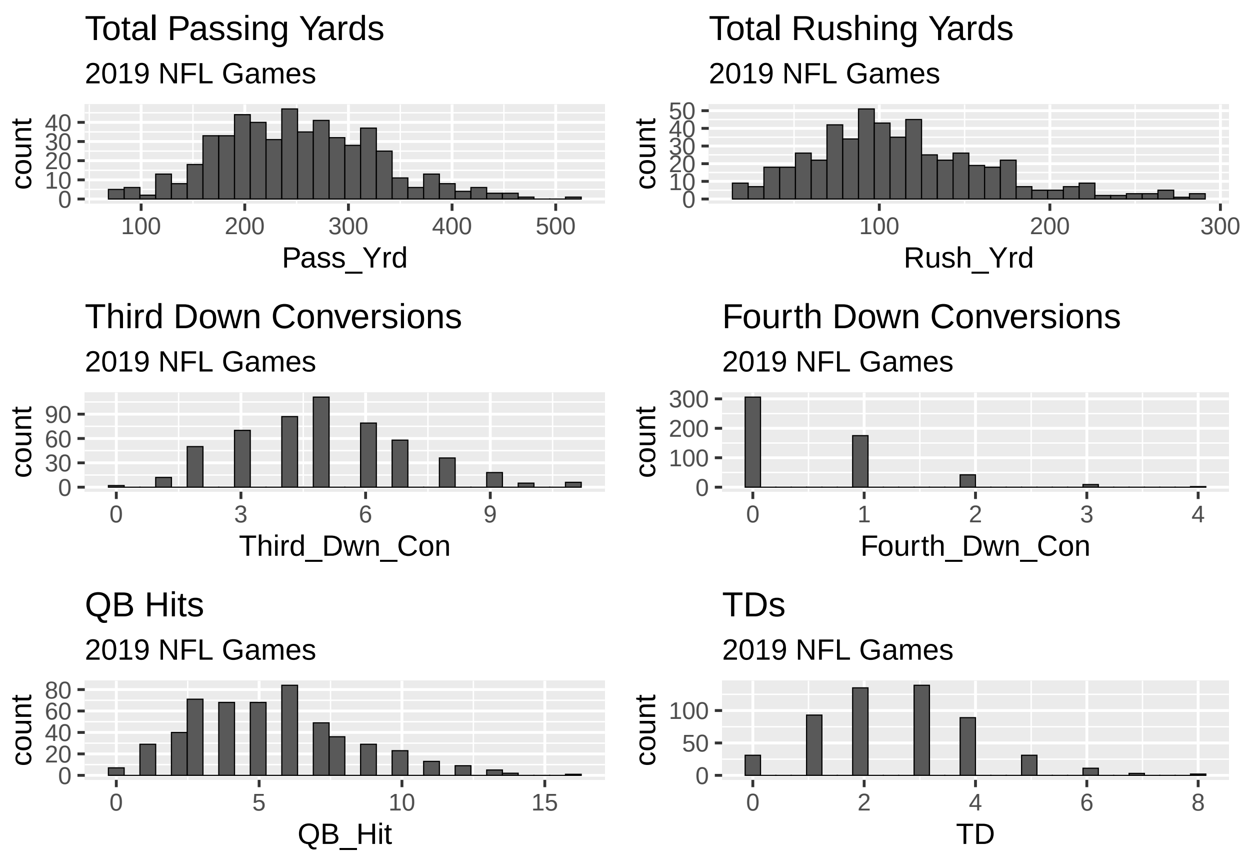
While we will have some variables with relatively clear-cut relationships with score differential, our analysis will also focus on how more "strategic" variables relate to success. More turnovers in a game and more total yards should correlate strongly with score differential, but we are interested in variables that have a more unclear relationship with the final outcome. Are missed extra point attempts indicative of the team as a whole having a bad day? What about third down conversion rate? And quarterback hits? These are the kinds of interactions and trends we hope to expose in our model. By using metrics that are not generally used to predict game outcomes, we hope to find insight into predicting wins that are not conventionally expected.

Data Cleaning, EDA:

Below, we graphed the distributions for a number of different potential predictors, as well as their relationship with the response variable of interest. Most predictors are slightly right-skewed, as they have a positive support and have a slight "bell" shape for the most common values but are stretched out by the few exceptional games where a team throws for 500 yards or rushes for 250 yards. However, it is mild right-skewness as it is not to the extent that log-transforming them would make the distributions more symmetric. For example, if we log-transform `Pass_Yrd` we actually find that the the resulting distribution is left-skewed, suggesting that the log transformation was actually too strong!

Furthermore, we can perform a quick check and see that the distribution of passing yards is roughly the same in 2019 as it was in 2018.


Total Passing Yards
2018 and 2019 NFL Games
count
0
20
40
100
200
300
400
500
Pass_Yrd
as.factor(year)
2018
2019



Initial Modeling:

The code for initializing the models has been hidden for space constraints. The linear model was using all predictors, the stepwise model stepped from an intercept only model to all predictors, and the randomforest has not been tuned (YET) and uses all predictors, as well as the default parameters. We can see that the RMSE's are pretty close together for all the models, but the stepwise linear model seems to outperform all.

```
##        rmse_2017 rmse_2018 rmse_2019 rmse_2020 rmse_2021
## Linear  9.992705  9.833446  9.919896  9.924871  10.14683
## Step    9.986854  9.799207  9.939556  9.923551  10.13224
## RF1    10.719757 10.323782 10.825754 10.519980  11.04696
```

Conclusion:

National Football League data provides a ripe opportunity for statistical analysis, especially when able to access play-by-play level data using the `nflfastR` package. While most variables have a slight right skew to them and only take on positive values, this did not present tremendous challenges for our models. Going forward, we're interested to continue investigating the impact of variables that provide a clear relationship towards "success" in a game–such as total passing yards or total rushing yards–as well as more "strategic" variables such as the number of third and fourth down conversions. Given the violations of symmetric distributions in many of the predictors and the potential for non-linear relationships to arise, examining the performance and interpretations of Random Forest models will also be a focus in our analysis.