NFL Play by Play Analysis, Linear, Logarithmic, and Random Forest Models

Brennan Frost and Conor Fogarty

2024-05-15

1.0 Project Description

For our project, we used play-by-play data across one if not more NFL seasons to try to see what "out-of-the-box" insights we could gain. Naturally, the domain of our project was sports, specifically the NFL (National Football League). The NFL is a professional American football league located in the United States that consists of 32 franchises that compete annually for the Super Bowl Championship. As aspiring data scientists, there are a number of different fields that we could end up working in. One of those fields, sports analytics, like a lot of analytical fields nowadays, is growing substantially. In the world of sports, it is becoming more and more valuable to generate mathematical and statistical insights that help respective franchises win games or matches. By learning how to apply techniques such as linear and logistic regression, or random forest modeling in the sports frame, we are making ourselves better suited in this field. For our project, we have 4 guiding questions that serve as our foundational thinking.

1. How does the specific formation of quarterback impact certain results?
2. How does the direction in which the running back rushed through impact certain results?
3. How does the type of pass impact certain results?
4. How does the accuracy of predicting a touchdown and/or predicting yardage change over a given time period?

We used linear regression to predict the yards gained or lost from both the rush direction and pass type. We used logistic regression to predict three different variables – whether the play resulted in a touchdown, whether the play carried out was a passing play, and whether the play carried out was a running play from the formation of the quarterback at the snap. We will also predict whether the play resulted in a touchdown from the rush direction, and whether the play resulted in a touchdown from the type of pass play. We used Random Forest modeling as another form of a predictive model to help evaluate whether a play results in a touchdown or not. We then took the results from those analyses and tried to observe whether the accuracy of the models (specifically the linear and random forest) increased, decreased, or remained stagnant over three seasons.

2.0 Data and Resources Used

Our data comes from NFLsavant.com, a website that houses play-by-play data from every game of every year for the past 10 NFL seasons. These are split up into 10 different .csv files. Each .csv file contains 45 different variables. Here are descriptions of the important variables that we used in some shape or form of the project;

Yards - int: The amount of yards gained or lost on the play Formation - factored with 7 levels: Where the quarterback was in relation to the football before the ball was snapped IsPass - factored with levels 0 and 1: Whether the play resulted in a pass IsRush - factored with levels 0 and 1: Whether the play resulted in a run IsTouchdown - factored with levels 0 and 1: Whether the play resulted in a touchdown PassType - factored with 6 levels: Where the quarterback throws the ball from the line of scrimmage relative to the field. RushDirection - factored with 7 levels: The player on the offensive line that the running back ran behind if the play was a run

PassType is where the quarterback throws the ball from the line of scrimmage relative to the field. The

dataset has six types of pass plays for the variable PassType; SHORT RIGHT, SHORT MIDDLE, SHORT LEFT, DEEP RIGHT, DEEP MIDDLE, and DEEP LEFT. For example, the NFL defines a deep pass 15 yards or more downfield from the line of scrimmage, so DEEP LEFT means that the quarterback threw the ball 15 yards or more down his left side of the field from the line of scrimmage. RushDirection is the player on the offensive line that the running back ran behind if the play was a run. The goal of a running back is to find holes in the offensive line to run through as they are smashing into the defensive line. These holes are created by players at the specific positions.The dataset has seven different positions for the variable Rush Direction; RIGHT TACKLE, RIGHT GUARD, RIGHT END, LEFT TACKLE, LEFT GUARD, LEFT END, and CENTER. Formation explains the specific formation of the quarterback relative to where the ball is on the line of the scrimmage. The dataset has six types of pass plays for the variable PassType; UNDER CENTER, SHOTGUN, NO HUDDLE SHOTGUN, NO HUDDLE, and PUNT. UNDER CENTER means that the quarterback squats behind the center (who is waiting to snap the football) ready to take the ball from the center on his command. SHOTGUN means that the quarterback is three to five yards behind the center waiting to catch the snap from the center. NO HUDDLE means that the offense is not huddling after a play has ended, wasting little time to get off the next play. NO HUDDLE SHOTGUN means the offense is running no huddle and the quarterback is in the SHOTGUN formation. This implies that NO HUDDLE also means that the quarterback is UNDER CENTER. PUNT is just the standard punt formation when the offensive team is getting ready to punt away the ball to the defensive team.

Here are the frequency tables of the important variables at play for our project.

Table 1: Figure 1: Frequency of Yards in 2023

Formation	Frequency of Plays
	636
FIELD GOAL	824
NO HUDDLE	381
NO HUDDLE SHOTGUN	2302
PUNT	1746
SHOTGUN	18038
UNDER CENTER	15545

Table 2: Figure 2: Frequency of Passing Plays in 2023

Pass	Frequency of Plays
0	24617
1	14855

Table 3: Figure 3: Frequency of Rushing Plays in 2023

Rush	Frequency of Plays
0	28526
1	10946

Table 4: Figure 4: Frequency of Touchdowns in 2023

Touchdown	Frequency of Plays
0	38414
1	1058

Touchdown	Frequency of Plays

Table 5: Figure 5: Frequency of Passing Types in 2023

Pass Type	Frequency of Plays
	24617
DEEP LEFT	1156
DEEP MIDDLE	480
DEEP RIGHT	1209
SHORT LEFT	4576
SHORT MIDDLE	2401
SHORT RIGHT	5033

Table 6: Figure 6: Frequency of Rush Directions in 2023

Rush Direction	Frequency of Plays
	29426
CENTER	2656
LEFT END	1287
LEFT GUARD	1284
LEFT TACKLE	1245
RIGHT END	1142
RIGHT GUARD	1351
RIGHT TACKLE	1081

Here is a histogram of the yardage distribution from the 2023 season. This plot shows the frequency in which a play resulted in a certain amount of yardage.

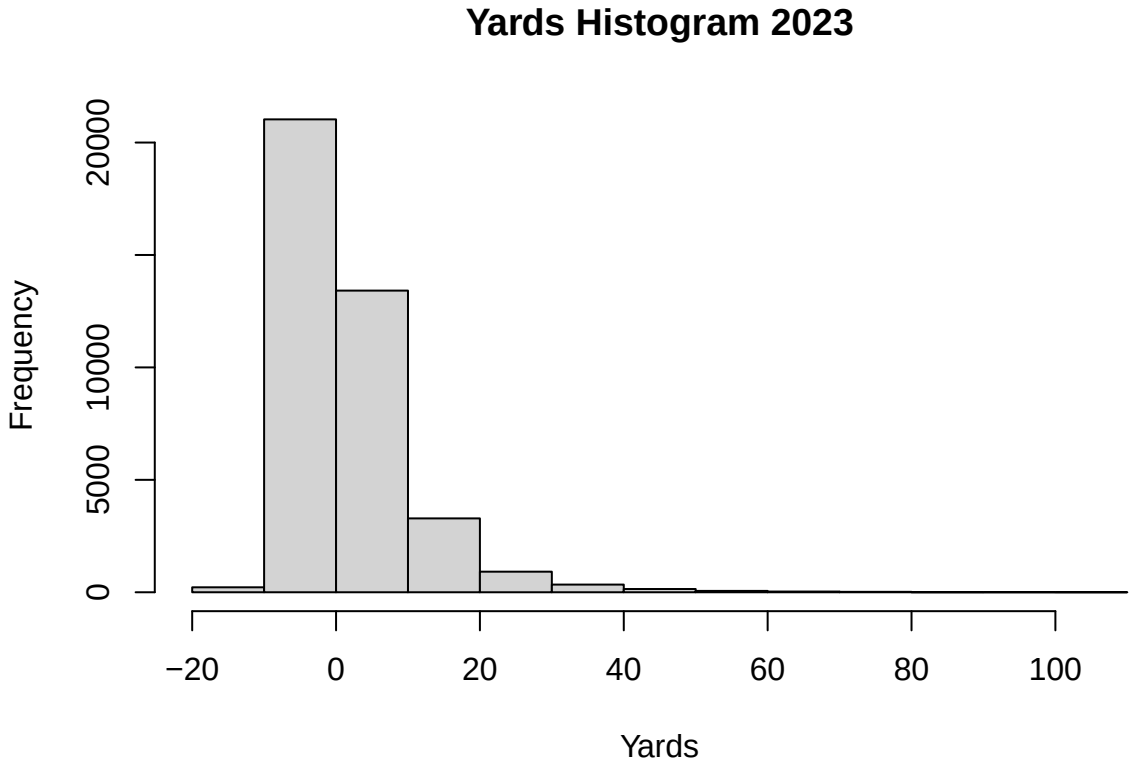

Figure 7: Distribution of Plays per Yards Gained

One thing to note is the data for each of the three seasons had to be cleaned in order for us to run the models. Since the eight models was prediciting from either formation, pass type, or rush direction, we made three different subsets from the overall data. One subset was plays from which one of the seven specific formations was present, excluding plays like kickoffs and punts. Another subset was plays from which one of the six specific passes were present. The last subset was plays from which one of the seven different rush directions were present.

3.0 Analysis

With the guiding questions in mind, we used three different modeling techniques to help us answer them; linear regression, logistic regression, and random forest modeling.

3.1 LINEAR REGRESSION We used linear regression to predict the yards gained or lost from both the rush direction and pass type. We ended up having to create a subset of just the pass plays from the existing dataframe, and another subset of just the rush plays from the existing dataframe. We then factored the independent variables – PassType and RushDirection – into levels. We also divided the data into training and testing sets. The training set is what each linear regression model will learn from, and the testing set is what each linear regression model will be evaluated on to see how well it performs. It does this division randomly, with 75% of the data used for training and 25% for testing.

We ran two linear regression models. One used Rush Direction to predict the total amount of yards gained (either positive or negative) on the play. Here are the results from the models. Each table corresponds to a different season.

Figure 8: Summary of Linear Model - Yards & Rush Direction (2023)

```
##
## Call:
## lm(formula = Yards ~ RushDirection, data = train.rush.data_2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -15.864  -3.021  -1.154   1.717  71.788
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.7404     0.1355  27.611  < 2e-16 ***
## RushDirectionLEFT END        1.1236     0.2365   4.751 2.06e-06 ***
## RushDirectionLEFT GUARD      0.2808     0.2379   1.180   0.2380
## RushDirectionLEFT TACKLE     0.4717     0.2366   1.994   0.0462 *
## RushDirectionRIGHT END       1.5427     0.2463   6.263 3.99e-10 ***
## RushDirectionRIGHT GUARD     0.4132     0.2323   1.779   0.0754 .
## RushDirectionRIGHT TACKLE    0.2475     0.2497   0.991   0.3218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.016 on 7529 degrees of freedom
## Multiple R-squared:  0.006937,   Adjusted R-squared:  0.006146
## F-statistic: 8.766 on 6 and 7529 DF,  p-value: 1.532e-09
```

Figure 9: Summary of Linear Model - Yards & Rush Direction (2022)

```
##
## Call:
## lm(formula = Yards ~ RushDirection, data = train.rush.data_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.301  -3.239  -1.239   1.556  76.806
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.9938     0.1391  28.712  < 2e-16 ***
## RushDirectionLEFT END        1.4505     0.2424   5.984 2.26e-09 ***
## RushDirectionLEFT GUARD      0.2833     0.2427   1.167  0.24318
## RushDirectionLEFT TACKLE     0.2002     0.2393   0.837  0.40278
## RushDirectionRIGHT END       1.3069     0.2524   5.178 2.29e-07 ***
## RushDirectionRIGHT GUARD     0.2455     0.2439   1.007  0.31414
## RushDirectionRIGHT TACKLE    0.7769     0.2439   3.185  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.345 on 8090 degrees of freedom
## Multiple R-squared:  0.007063,   Adjusted R-squared:  0.006327
## F-statistic: 9.592 on 6 and 8090 DF,  p-value: 1.544e-10
```

Figure 10: Summary of Linear Model - Yards & Rush Direction (2021)

```
##
## Call:
## lm(formula = Yards ~ RushDirection, data = train.rush.data_2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.042  -4.042  -1.395   1.728  86.958
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.04180    0.13172  30.686  < 2e-16 ***
## RushDirectionCENTER        -0.18597    0.18995  -0.979    0.328
## RushDirectionLEFT END       1.04029    0.23922   4.349 1.38e-05 ***
```

```
## RushDirectionLEFT GUARD    0.03304    0.23642   0.140    0.889
## RushDirectionLEFT TACKLE   0.35324    0.23497   1.503    0.133
## RushDirectionRIGHT END     1.23027    0.24746   4.972 6.74e-07 ***
## RushDirectionRIGHT GUARD  -0.10093    0.23234  -0.434    0.664
## RushDirectionRIGHT TACKLE  0.30648    0.24112   1.271    0.204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.538 on 11237 degrees of freedom
## Multiple R-squared:  0.004966,   Adjusted R-squared:  0.004346
## F-statistic: 8.012 on 7 and 11237 DF,  p-value: 9.643e-10
```

Another linear regression model used PassType to predict the total amount of yards gained (either positive or negative) on the play. Here are the results from the models. Each table corresponds to a different season.

Figure 11: Summary of Linear Model - Yards & PassType (2023)

```
##
## Call:
## lm(formula = Yards ~ PassType, data = train.pass.data_2023)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.066  -5.903  -1.740   4.260  93.097
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          11.066202   0.327791  33.760  < 2e-16 ***
## PassTypeDEEP MIDDLE   2.332674   0.606063   3.849 0.000119 ***
## PassTypeDEEP RIGHT   -0.007126   0.454770  -0.016 0.987499
## PassTypeSHORT LEFT   -5.162719   0.367087 -14.064  < 2e-16 ***
## PassTypeSHORT MIDDLE -3.633292   0.398159  -9.125  < 2e-16 ***
## PassTypeSHORT RIGHT  -5.326213   0.363068 -14.670  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.618 on 11137 degrees of freedom
## Multiple R-squared:  0.05005,    Adjusted R-squared:  0.04963
## F-statistic: 117.4 on 5 and 11137 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of Linear Model - Yards & PassType (2022)

```
##
## Call:
## lm(formula = Yards ~ PassType, data = train.pass.data_2022)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.116  -6.099  -1.614   4.386  93.239
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.9700     0.3261  33.641  < 2e-16 ***
## PassTypeDEEP MIDDLE    3.1463     0.5586   5.632 1.82e-08 ***
## PassTypeDEEP RIGHT    -0.1294     0.4575  -0.283    0.777
## PassTypeSHORT LEFT    -4.8711     0.3632 -13.412  < 2e-16 ***
## PassTypeSHORT MIDDLE  -3.3560     0.3856  -8.703  < 2e-16 ***
## PassTypeSHORT RIGHT   -5.2088     0.3594 -14.494  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.406 on 11540 degrees of freedom
## Multiple R-squared:  0.05227,    Adjusted R-squared:  0.05186
## F-statistic: 127.3 on 5 and 11540 DF,  p-value: < 2.2e-16
```

Figure 13: Summary of Linear Model - Yards & PassType (2021)

```
##
## Call:
## lm(formula = Yards ~ PassType, data = train.pass.data_2021)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.823  -5.831  -1.823   3.842  80.630
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.8229     0.2300   7.924 2.47e-15 ***
## PassTypeBACK TO           0.1771     9.0511   0.020    0.984
## PassTypeDEEP LEFT         8.5458     0.3761  22.724  < 2e-16 ***
## PassTypeDEEP MIDDLE      13.5603     0.4942  27.440  < 2e-16 ***
## PassTypeDEEP RIGHT        8.5470     0.3677  23.244  < 2e-16 ***
## PassTypeLEFT TO           6.1771     9.0511   0.682    0.495
## PassTypeNOT LISTED       -6.8229     9.0511  -0.754    0.451
## PassTypeRIGHT. PENALTY   -1.8229     9.0511  -0.201    0.840
## PassTypeSHORT LEFT        4.3349     0.2710  15.998  < 2e-16 ***
## PassTypeSHORT MIDDLE      5.6572     0.2940  19.240  < 2e-16 ***
## PassTypeSHORT RIGHT       4.0079     0.2678  14.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.048 on 14684 degrees of freedom
## Multiple R-squared:  0.07752,    Adjusted R-squared:  0.07689
## F-statistic: 123.4 on 10 and 14684 DF,  p-value: < 2.2e-16
```

In 2021, we experienced significantly different types of predictors then what we will see in the datasets from 2022 and 2023, however it is still decipherable, with roughly an intercept of 1.82 yards, with deep left, middle, and right being significant predictors in yardage gains between 8.5-13.56 yards. While we also saw short left, middle, and right being significant predictors of 4-5.65 yards gained.

In 2022, we saw an intercept of roughly 10.97 yards, with the significant predictors being a pass type of deep middle, averaging 3.15 yards over the intercept with a p-value of 0.000119. We also saw short left, middle, and right being significant predictors of successful passes however they reduced the intercept by roughly 3.5-5.2 yards.

In 2023, we saw an intercept of roughly 11.07 yards, with the significant predictors being a pass type of deep middle, averaging 2.33 yards over the intercept with a p-value of 0.000119. We also saw short left, middle, and right being significant predictors of successful passes however they reduced the intercept by roughly 3.6-5.3 yards

In the overall trends, we saw the ability of deep passes to gain high yardage (14 yards or more)decrease throughout the three seasons, with 2023 losing also a full yard per deep pass gain compared to 2021 and 2022. The short pass yard gainage stayed relatively staggent throughout the three seasons, averaging between 5.77 to 7.47, with a slightly higher average in 2022 and 2023, however that could be attributed to the change of data collection.

3.2 LOGISTIC REGRESSION We used logistic regression for three processes. One was to predict three different variables – whether the play resulted in a touchdown (isTouchdown), whether the play carried out was a passing play (isPass), and whether the play carried out was a running play (isRush) from the formation of the quarterback at the snap (Formation). Another was to predict whether the play resulted in a touchdown (isTouchdown) from the rush direction (RushDirection). The last process was to predict whether the play resulted in a touchdown (isTouchdown) from the type of pass play (PassType). We had to create an additional subset of just plays that have a specific quarterback formation from the existing dataframe. We then factored the independent variable – Formation – into levels. Like the linear regression models, we divided the data into training and testing sets.

We then ran five logistic regression models. One used Formation to predict whether the play carried out ended up as a touchdown or not. Here are the results from the models. Each table corresponds to a different season.

Figure 14: Summary of Logistic Model - IsTouchdown ~ Formation (2023)

```
##
## Call:
## glm(formula = IsTouchdown ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2023)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -6.4184     0.9994  -6.422 1.34e-10 ***
## FormationNO HUDDLE           3.5060     1.0339   3.391 0.000697 ***
## FormationNO HUDDLE SHOTGUN   2.7861     1.0107   2.757 0.005842 **
## FormationPUNT                0.8334     1.0952   0.761 0.446705
## FormationSHOTGUN             3.1799     1.0004   3.179 0.001480 **
## FormationUNDER CENTER        2.4536     1.0017   2.449 0.014310 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7279.6  on 29127  degrees of freedom
## Residual deviance: 7117.3  on 29122  degrees of freedom
## AIC: 7129.3
##
## Number of Fisher Scoring iterations: 8
```

Figure 15: Summary of Logistic Model - IsTouchdown ~ Formation (2022)

```
##
## Call:
## glm(formula = IsTouchdown ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2022)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -3.4965     0.2621 -13.341  < 2e-16 ***
## FormationNO HUDDLE SHOTGUN   0.1135     0.2926   0.388    0.698
## FormationPUNT               -2.6179     0.6345  -4.126 3.69e-05 ***
## FormationSHOTGUN             0.2789     0.2661   1.048    0.294
## FormationUNDER CENTER       -0.1701     0.2691  -0.632    0.527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 7585.9  on 27642  degrees of freedom
## Residual deviance: 7485.2  on 27638  degrees of freedom
## AIC: 7495.2
##
## Number of Fisher Scoring iterations: 8
```

Figure 16: Summary of Logistic Model - IsTouchdown ~ Formation (2021)

```
##
## Call:
## glm(formula = IsTouchdown ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2021)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.44255    0.26230 -13.125  < 2e-16 ***
## FormationNO HUDDLE SHOTGUN  0.07765    0.28839   0.269    0.788
## FormationPUNT              -3.10780    0.75295  -4.127 3.67e-05 ***
## FormationSHOTGUN            0.33806    0.26563   1.273    0.203
## FormationUNDER CENTER      -0.22406    0.26835  -0.835    0.404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8827.6  on 30676  degrees of freedom
## Residual deviance: 8682.7  on 30672  degrees of freedom
## AIC: 8692.7
##
## Number of Fisher Scoring iterations: 8
```

Another used Formation to predict whether the play carried out ended up as a running play or not. Here are the results from the models. Each table corresponds to a different season.

Figure 17: Summary of Logistic Model - IsRush ~ Formation (2023)

```
##
## Call:
## glm(formula = IsRush ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2022)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.47957    0.09112   5.263 1.42e-07 ***
## FormationNO HUDDLE SHOTGUN  -1.59114    0.10579 -15.040  < 2e-16 ***
## FormationPUNT              -17.04564   65.06686  -0.262    0.793
## FormationSHOTGUN            -1.39767    0.09320 -14.997  < 2e-16 ***
## FormationUNDER CENTER       -0.92502    0.09317  -9.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34537  on 27642  degrees of freedom
## Residual deviance: 32936  on 27638  degrees of freedom
## AIC: 32946
##
## Number of Fisher Scoring iterations: 15
##
```

Figure 18: Summary of Logistic Model - IsRush ~ Formation (2022)

```
## Call:
## glm(formula = IsRush ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2022)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.47957    0.09112   5.263 1.42e-07 ***
## FormationNO HUDDLE SHOTGUN  -1.59114    0.10579 -15.040  < 2e-16 ***
## FormationPUNT              -17.04564   65.06686  -0.262    0.793
## FormationSHOTGUN            -1.39767    0.09320 -14.997  < 2e-16 ***
## FormationUNDER CENTER       -0.92502    0.09317  -9.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34537  on 27642  degrees of freedom
## Residual deviance: 32936  on 27638  degrees of freedom
## AIC: 32946
##
## Number of Fisher Scoring iterations: 15
```

Figure 19: Summary of Logistic Model - IsRush ~ Formation (2021)

```
## Call:
## glm(formula = IsRush ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2021)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.29130    0.09188   3.171  0.00152 **
## FormationNO HUDDLE SHOTGUN  -1.47011    0.10498 -14.004  < 2e-16 ***
## FormationPUNT              -16.85737   64.10770  -0.263  0.79259
## FormationSHOTGUN            -1.35998    0.09392 -14.480  < 2e-16 ***
## FormationUNDER CENTER       -0.71608    0.09364  -7.647 2.05e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37777  on 30676  degrees of freedom
## Residual deviance: 35923  on 30672  degrees of freedom
## AIC: 35933
##
## Number of Fisher Scoring iterations: 15
```

The third model used Formation to predict whether the play carried out ended up as a passing play or not. Here are the results from the models. Each table corresponds to a different season.

Figure 20: Summary of Logistic Model - IsPass ~ Formation (2023)

```
##
## Call:
## glm(formula = IsPass ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2023)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                -1.757e+01  1.597e+02  -0.110    0.912
## FormationNO HUDDLE          1.638e+01  1.597e+02   0.103    0.918
## FormationNO HUDDLE SHOTGUN  1.821e+01  1.597e+02   0.114    0.909
## FormationPUNT              -2.487e-09  1.929e+02   0.000    1.000
## FormationSHOTGUN            1.809e+01  1.597e+02   0.113    0.910
## FormationUNDER CENTER       1.559e+01  1.597e+02   0.098    0.922
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38732  on 29127  degrees of freedom
## Residual deviance: 29025  on 29122  degrees of freedom
## AIC: 29037
##
## Number of Fisher Scoring iterations: 16
```

Figure 21: Summary of Logistic Model - IsPass ~ Formation (2022)

```
##
## Call:
## glm(formula = IsPass ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2022)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.95226    0.09879  -9.639  < 2e-16 ***
## FormationNO HUDDLE SHOTGUN   1.75484    0.11081  15.837  < 2e-16 ***
## FormationPUNT              -16.61381  107.27705  -0.155    0.877
## FormationSHOTGUN             1.51635    0.10049  15.090  < 2e-16 ***
## FormationUNDER CENTER       -0.62460    0.10195  -6.127 8.98e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37530  on 27642  degrees of freedom
## Residual deviance: 29846  on 27638  degrees of freedom
## AIC: 29856
##
## Number of Fisher Scoring iterations: 16
```

Figure 22: Summary of Logistic Model - IsPass ~ Formation (2021)

```
##
## Call:
## glm(formula = IsPass ~ Formation, family = binomial(link = "logit"),
##     data = train.form.data_2021)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.75263    0.09742  -7.725 1.11e-14 ***
## FormationNO HUDDLE SHOTGUN  1.64790    0.10837  15.206  < 2e-16 ***
## FormationPUNT              -6.49159    1.00446  -6.463 1.03e-10 ***
## FormationSHOTGUN            1.44221    0.09908  14.557  < 2e-16 ***
## FormationUNDER CENTER      -0.80265    0.10017  -8.013 1.12e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 41890  on 30676  degrees of freedom
## Residual deviance: 32694  on 30672  degrees of freedom
## AIC: 32704
##
## Number of Fisher Scoring iterations: 9
```

The fourth model used RushDirection to predict whether the play carried out ended up as a touchdown or not. Here are the results from the models. Each table corresponds to a different season.

Figure 23: Summary of Logistic Model - IsTouchdown ~ RushDirection (2023)

```
##
## Call:
## glm(formula = IsTouchdown ~ RushDirection, family = binomial(link = "logit"),
##     data = train.rush.data_2023)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.0172     0.1068 -28.257  < 2e-16 ***
## RushDirectionLEFT END      -0.2027     0.1988  -1.020  0.30779
## RushDirectionLEFT GUARD    -0.4016     0.2141  -1.876  0.06067 .
## RushDirectionLEFT TACKLE   -0.6485     0.2327  -2.787  0.00532 **
## RushDirectionRIGHT END     -0.1073     0.2010  -0.534  0.59336
## RushDirectionRIGHT GUARD   -0.6626     0.2289  -2.894  0.00380 **
## RushDirectionRIGHT TACKLE  -0.6741     0.2503  -2.693  0.00707 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2288.2  on 7535  degrees of freedom
## Residual deviance: 2269.3  on 7529  degrees of freedom
## AIC: 2283.3
##
## Number of Fisher Scoring iterations: 6
```

Figure 24: Summary of Logistic Model - IsTouchdown ~ RushDirection (2022)

```
##
## Call:
## glm(formula = IsTouchdown ~ RushDirection, family = binomial(link = "logit"),
##     data = train.rush.data_2022)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.89546    0.09839 -29.427  < 2e-16 ***
## RushDirectionLEFT END     -0.47387    0.20026  -2.366  0.01797 *
## RushDirectionLEFT GUARD   -0.38220    0.19424  -1.968  0.04910 *
## RushDirectionLEFT TACKLE  -0.37148    0.19053  -1.950  0.05121 .
## RushDirectionRIGHT END    -0.41417    0.20512  -2.019  0.04347 *
## RushDirectionRIGHT GUARD  -0.61866    0.21258  -2.910  0.00361 **
## RushDirectionRIGHT TACKLE -0.39522    0.19620  -2.014  0.04397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2669.4  on 8096  degrees of freedom
## Residual deviance: 2655.5  on 8090  degrees of freedom
```

```
## AIC: 2669.5
##
## Number of Fisher Scoring iterations: 6
```

Figure 25: Summary of Logistic Model - IsTouchdown ~ RushDirection (2021)

```
##
## Call:
## glm(formula = IsTouchdown ~ RushDirection, family = binomial(link = "logit"),
##     data = train.rush.data_2021)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.51762    0.12043 -29.210   <2e-16 ***
## RushDirectionCENTER        0.49582    0.15619   3.174   0.0015 **
## RushDirectionLEFT END      0.45735    0.19056   2.400   0.0164 *
## RushDirectionLEFT GUARD    0.46457    0.18827   2.468   0.0136 *
## RushDirectionLEFT TACKLE   0.28857    0.19667   1.467   0.1423
## RushDirectionRIGHT END     0.10254    0.21868   0.469   0.6391
## RushDirectionRIGHT GUARD   0.04123    0.20966   0.197   0.8441
## RushDirectionRIGHT TACKLE  0.02700    0.21850   0.124   0.9017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3553.0  on 11244  degrees of freedom
## Residual deviance: 3534.8  on 11237  degrees of freedom
## AIC: 3550.8
##
## Number of Fisher Scoring iterations: 6
```

The last model used PassType to predict whether the play carried out ended up as a touchdown or not. Here are the results from the models. Each table corresponds to a different season.

Figure 26: Summary of Logistic Model - IsTouchdown ~ PassType (2023)

```
##
## Call:
## glm(formula = IsTouchdown ~ PassType, family = binomial(link = "logit"),
##     data = train.pass.data_2023)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.89775    0.15313 -18.924   <2e-16 ***
## PassTypeDEEP MIDDLE    0.22510    0.26444   0.851   0.3946
## PassTypeDEEP RIGHT     0.12972    0.20681   0.627   0.5305
## PassTypeSHORT LEFT    -0.26899    0.17622  -1.526   0.1269
## PassTypeSHORT MIDDLE -0.08874    0.18847  -0.471   0.6377
## PassTypeSHORT RIGHT   -0.44956    0.17740  -2.534   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3933.2  on 11142  degrees of freedom
## Residual deviance: 3913.0  on 11137  degrees of freedom
## AIC: 3925
##
```

```
## Number of Fisher Scoring iterations: 6
```

Figure 27: Summary of Logistic Model - IsTouchdown ~ PassType (2022)

```
##
## Call:
## glm(formula = IsTouchdown ~ PassType, family = binomial(link = "logit"),
##     data = train.pass.data_2022)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.83832    0.15170 -18.711   <2e-16 ***
## PassTypeDEEP MIDDLE   0.47542    0.22931   2.073   0.0381 *
## PassTypeDEEP RIGHT    0.07558    0.20942   0.361   0.7182
## PassTypeSHORT LEFT   -0.42098    0.17642  -2.386   0.0170 *
## PassTypeSHORT MIDDLE -0.32143    0.18782  -1.711   0.0870 .
## PassTypeSHORT RIGHT  -0.31372    0.17196  -1.824   0.0681 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4154.6  on 11545  degrees of freedom
## Residual deviance: 4128.0  on 11540  degrees of freedom
## AIC: 4140
##
## Number of Fisher Scoring iterations: 6
```

Figure 28: Summary of Logistic Model - IsTouchdown ~ PassType (2021)

```
##
## Call:
## glm(formula = IsTouchdown ~ PassType, family = binomial(link = "logit"),
##     data = train.pass.data_2021)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -4.1504     0.2057 -20.174  < 2e-16 ***
## PassTypeBACK TO          -8.4157   324.7438  -0.026    0.979
## PassTypeDEEP LEFT         1.2882     0.2519   5.113 3.16e-07 ***
## PassTypeDEEP MIDDLE       1.6347     0.2759   5.926 3.11e-09 ***
## PassTypeDEEP RIGHT        1.4391     0.2440   5.899 3.67e-09 ***
## PassTypeLEFT TO          -8.4157   324.7438  -0.026    0.979
## PassTypeNOT LISTED       -8.4157   324.7438  -0.026    0.979
## PassTypeRIGHT. PENALTY   -8.4157   324.7438  -0.026    0.979
## PassTypeSHORT LEFT        0.9804     0.2209   4.438 9.07e-06 ***
## PassTypeSHORT MIDDLE      1.2725     0.2246   5.665 1.47e-08 ***
## PassTypeSHORT RIGHT       0.9934     0.2195   4.526 6.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5244.4  on 14694  degrees of freedom
## Residual deviance: 5180.6  on 14684  degrees of freedom
## AIC: 5202.6
##
## Number of Fisher Scoring iterations: 11
```

Response Variable: IsTouchdown, Predictor Variable: Formation In 2021 and 2022, we were unable to identify any significant predictors using logarithmic regression to determine if formations dictate touchdown likelihood, however in 2023 we were able to identify significant predictors, which we ordered in from highest to lowest p-value in this order respectively, No Huddle(p-value: 0.000697), Shotgun (p-value: 0.001480), No huddle Shotgun (p-value: 0.005842), and under center (p-value: 0.014310). We believe that the issue with 2021 and 2022 was the form of data collection done by NFLSavant, however we were not able to specifically locate the issue. We also attempted to use ROSE synthetic data, however this resulted in all three years being unable to identify significant predictors.

Response Variable: IsRush, Predictor Variable: Formation In 2021, we were able to identify this predictors of IsRush within variable Formation in increasing p-value order, Shotgun and NoHuddleShotgun were tied (p-values: <2e-16), Under Center (p-value 2.05e-14), and Huddle Shotgun (p-value: 2.05e-14). In 2022, we were able to identify this predictors of IsRush within variable Formation in increasing p-value order, Shotgun, Under Center, and NoHuddleShotgun were tied (p-values: <2e-16), and Huddle Shotgun (p-value: 1.42e-07). However, in 2023, our model was unable to identify any predictors, we believe that this was due to an issue with the way estimate/intercept was calculated using data from the 2023, which had changed its recording methodology that year, we were unfortunately not able to resolve this issue.

Response Variable: IsPass, Predictor Variable: Formation In 2021, we were able to identify this predictors of IsPass within variable Formation in increasing p-value order, Shotgun and NoHuddleShotgun were tied (p-values: <2e-16), and Under Center (p-value 1.12e-15). In 2022, we were able to identify this predictors of IsPass within variable Formation in increasing p-value order, Shotgun,and NoHuddleShotgun were tied (p-values: <2e-16), and Under Center (p-value: 8.98e-10).However, in 2023, our model was unable to identify any predictors, we believe that this was due to the same issue we faced in our 2023 IsRush logarithmic regression model, and we were again unable to identify why the model was unsuccessful.

Response Variable: IsTouchdown, Predictor Variable: RushDirection In 2021, we were able to identify this predictors of IsTouchdown within variable RushDirection in increasing p-value order, Center(p-value: 0.0015), and Left End(p-value: 0.0136), and Left End (p-value: 0.164). In 2022, we were able to identify this predictors of IsTouchdown within variable RushDirection in increasing p-value order, Right Guard(p-value: 0.00361), Left End(p-value: 0.01797), Right End(p-value: 0.04347), and Right Tackle(p-value: 0.04397). In 2023, we were able to identify this predictors of IsTouchdown within variable RushDirection in increasing p-value order, Right Guard(p-value: 0.00380) , Left Tackle(p-value: 0.00532), and Right Tackle(p-value: 0.00707).

Response Variable: IsTouchdown, Predictor Variable: PassType In 2021, we were able to identify this predictors of IsTouchdown within variable PassType in increasing p-value order, Short Right(p-value: 6.02e-06), Short Left(p-value: 9.07e-06), Deep Left(p-value: 3.16e-07), Short Middle(p-value: 1.47e-08), Deep Middle(p-value: 3.11e-09) and Deep Right (p-value: 3.67e-09). In 2022, we were able to identify this predictors of IsTouchdown within variable PassType in increasing p-value order, Short Left (p-value: 0.0170) and Deep Middle (p-value: 0.0381). In 2023, we were able to identify these predictors of IsTouchdown within the PassType, however we were only able to identify one significant predictor variable, Short Right(p-value: 0.0113).

3.3 RANDOM FOREST MODELING For our third modeling tool, we built and evaluated three Random Forest models. The model was built to predict whether a play results in a touchdown or not. One model evaluated the 2023 season, one model evaluated the 2022 season, and the last model evaluated the 2021 season. The model was trained on the data after applying an oversampling technique. Once the models were trained, it used them to make predictions on the test set, which contains data the models haven't seen before. Then, it evaluated the performance of the models using confusion matrices.

Here are the results from the models. Each confusion matrix corresponds to a different season.

Figure 29: Confusion Matrix Summary - IsTouchdown (2023)

```
##            Reference
## Prediction    0    1
##          0 8361   72
##          1 1083  192
```

```
##            Reference
## Prediction    0    1
##          0 7760   74
##          1 1170  209
```

Figure 30: Confusion Matrix Summary - IsTouchdown (2022)

```
##            Reference
## Prediction    0    1
##          0 8617   97
##          1 1275  236
```

Figure 31: Confusion Matrix Summary - IsTouchdown (2022)

Randon Forest Description

In 2021, our random forest model correctly predicted if a play was a touchdown 98.89% of the time, while it predicted a non-touchdown 15. 77%, with a total accuracy of 86.68%.

In 2022, our random forest model correctly predicted if a play was a touchdown 99.05% of the time, while it predicted a non-touchdown 15.16%, with a total accuracy of 86.49%.

In 2023, our random forest model correctly predicted if a play was a touchdown 99.14% of the time, while it predicted a non-touchdown 15.05%, with a total accuracy of 88.10%.

All Random Forest Tree models were built around synthetic data, created using ROSE and then tested against real data. As you can see above, correctly predicting touchdown hovered around 99%, with correctly predicting non-touchdown stayed around 15 and total accuracy was roughly between 86% and 88%. We hoped that the synthetic data would help to correctly predict the negative outcome (non-touchdown), however that does not seem to be the case.

4.0 Conclusions

Linear Regression Analysis From the linear regression analysis of pass types over the three years (2021-2023), several key trends and shifts have been identified:

Deep Passes: Deep left, middle, and right passes were significant predictors of higher yardage gains in 2021, with gains ranging from 8.5 to 13.56 yards. In 2022 and 2023, deep middle passes remained significant but with reduced yardage gains over the intercept (3.15 yards in 2022 and 2.33 yards in 2023). This indicates a decline in the effectiveness of deep passes over the period, with 2023 showing the lowest additional gain per deep pass. Short Passes:

Short Passes: Short passes (left, middle, and right) consistently remained significant predictors of yardage gains across all three years. However, they reduced the intercept yardage by 3.5-5.2 yards in 2022 and 3.6-5.3 yards in 2023, compared to 4-5.65 yards in 2021.

Overall Yardage Trends: The ability of deep passes to gain high yardage (14 yards or more) decreased over the three seasons, with a notable drop of a full yard in 2023 compared to previous years. Short pass yardage remained relatively stable, averaging between 5.77 to 7.47 yards, with slightly higher averages in 2022 and 2023.

Logarithmic Regression Analysis The analysis of touchdown likelihood and rush/pass likelihood based on formations and rush directions provided mixed results:

Touchdown Likelihood (IsTouchdown): In 2021 and 2022, no significant predictors were identified for formations affecting touchdown likelihood. In 2023, significant predictors emerged for formations, with No Huddle, Shotgun, No Huddle Shotgun, and Under Center formations showing varying levels of significance, suggesting improvements in data collection or changes in play strategies.

Rush Likelihood (IsRush): Significant predictors were identified in 2021 and 2022 for formation types influencing rush likelihood, with Shotgun and No Huddle Shotgun being the most significant. In 2023, no significant predictors were identified, likely due to issues with data recording methodologies.

Pass Likelihood (IsPass):Significant predictors were identified in 2021 and 2022 for formations influencing pass likelihood, with Shotgun and No Huddle Shotgun being the most significant. Similar to rush likelihood, 2023 did not yield significant predictors, pointing to potential data issues.

Rush Direction and Touchdown Likelihood:Significant predictors for rush direction affecting touchdown likelihood were identified each year, with different rush directions being significant each season. This indicates variability in effective rush strategies over the years.

Pass Type and Touchdown Likelihood:Significant predictors for pass types affecting touchdown likelihood were identified each year, but the specific types varied. In 2021, multiple pass types were significant, while in 2022 and 2023, fewer pass types were significant.

Random Forest Models The random forest models showed high accuracy in predicting touchdowns across all three years, with the following observations:

Touchdown Prediction Accuracy: The models correctly predicted touchdowns around 99% of the time each year, showing high reliability in positive outcome prediction.

Non-Touchdown Prediction Accuracy: The accuracy for predicting non-touchdowns remained low, around 15%, despite using synthetic data generated by ROSE. This indicates a limitation in the model's ability to identify non-touchdown plays effectively.

Overall Accuracy: The total accuracy of the models ranged from 86.49% to 88.10%, showing consistent performance but highlighting the challenge of improving non-touchdown prediction accuracy.

Summary The analysis across linear regression, logarithmic regression, and random forest models reveals several insights into the effectiveness of different play types and formations over the years. Significant changes in data collection or play strategies are evident, especially in 2023. While the models perform well in predicting touchdowns, there is room for improvement in predicting non-touchdown outcomes. The variability in significant predictors across years underscores the dynamic nature of play strategies and data recording practices in football.

5.0 References

https://nflsavant.com/about.php