When do NFL Teams Run The Ball?

STAT835: Final Project, 2020

Brenton Summers

Department of Biostatistics and Data Science
University of Kansas, USA
December 13, 2020

Contents

Abstract

Understanding and predicting when a National Football Team (NFL) will run or pass allows the defending team to better strategize for the current play. This knowledge can lead to improved defensive performance and more wins. Finding the right factors allows a defensive coordinator to understand the scenario in real time and adjust his defense accordingly to counter the opposing offense.

The objective of this study is to find accurate and valuable predictor variables which influence the play type of a given play. The dataset used from this study is scraped NFL play data made available by the CRAN package, NFLFastR and a public github.

The variables collected included the down, yards to go, quarter, win probability (WP), running back (RB) salary, and field position. Win Probability is a common metric with multiple popular models. For this report we will use the given WP model from NFLFastR. RB salary will be generated from the projected starter at the beginning of the season.

Introduction

In the NFL, play type is a binary variable that can either be a run or a pass. Run plays consist of designed plays in which the football is given to a RB without the ball traveling in the air or when there is a designed quarterback (QB) run. Pass plays are any play in which the QB throws the ball in attempt to gain yards. For this exercise, QB scrambles have been omitted as these are initially pass plays that break down and then turn into a run play.

Understanding when the opposing team is more likely to run or pass allows a team to adjust their defense in real time to better disrupt the offense. A team that understands the opposition's tendencies will always have an advantage in any competition. Likewise, a team that is more predictable can then be easily countered. Understanding your own tendencies and shortcomings is a good way of self-reflection and evaluation.

Football teams can anticipate what play type is coming by evaluating the game setting including down, yards to go, quarter, WP, RB salary, and field position. Understanding the cause and effect relationship between play type and the predictor variables can yield to improved defensive performance and more team wins.

Primary Analysis Objectives

To investigate the association between the play type, down, yards to go, quarter, WP,RB salary, and field position as well as to predict the play type using the same variables.

Materials and Methods

Data Sources

The dataset was obtained online via public github for NFLFastR. Here is GitHub.

Variables in the dataset include **down**, **yards to go**, **quarter**, **WP**, **RB Salary**, and **field position**. The down and yards to go are the goals for the current play. The quarter is what part of the game the play occurred in. WP serves as a proxy for point differential and difference in team strength. It is a common theme that a winning team will call run plays more often while losing teams will call pass plays more often. RB salary is the amount of money committed to the team's starting RB at the beginning of the year.

This dataset was chosen based off interest in NFL data and a curiousity in the predictability of play type.

Statistical Analysis

The data is available in .csv (Comma Seperated Values) format. The data analysis is done using the statistical software R version 3.6.1 (2019-07-10) and the project focuses mainly on general linear models using the logit link. Each of the predictor variable is explored individually and the illustrations used are conducted on the entire dataset for the preliminary investigation. No missing values were found in the dataset. The large sample

size and absence of missing value is assumed to ensure better predictibility and less sampling variability. Automatic model selection method has been used to arrive at the final model. The model assumptions are assessed and a final model is decided upon.

Model Assumptions

All inferences are conducted using alpha = 0.05 unless stated otherwise. No adjustments for multiplicity are made as this is an exploratory analysis. Discrete variables are summarized with proportions and frequencies. Continuous variables are summarized using the basic statistics.

Primary Objective Analysis

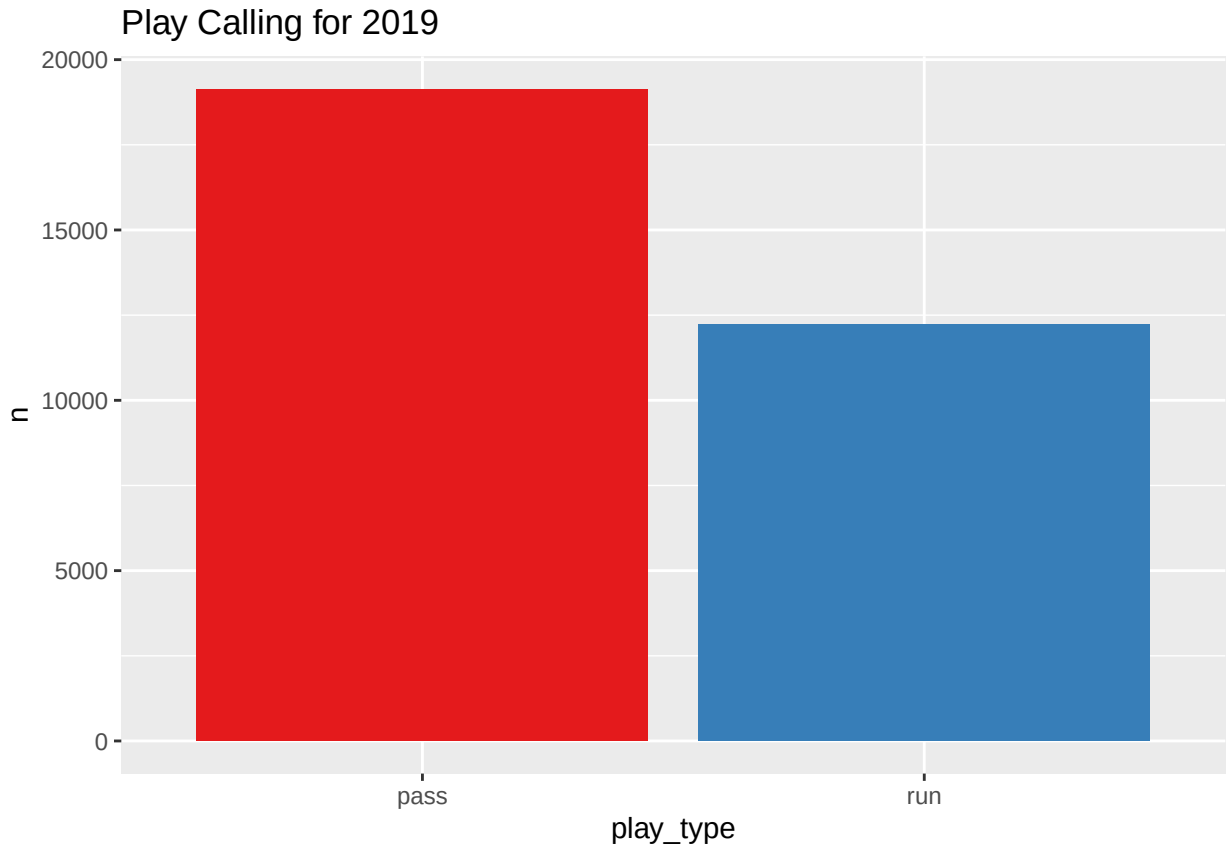
Exploring individual predictors and the response variable is very important before starting data analysis. It helps in detecting skewness, presence of outliers or can also suggest if transformations are necessary to fit a better model. Once the dataset variables are explored, the next step is performed to check the linearity between the predictors and the outcome individually. This helps us fit a better model which can do a better job at explaining the variation of the response.

Analysis of Play Type

Preliminary data analysis on the play type distribution shows that we have ample instances of both pass plays and run plays. In 2019 NFL teams ran the ball on approximately 39% of plays (12,235 plays) and passed the ball on approximately 61% of plays (19,138 plays). There is no significant skewness or need to transform the response variable in this analysis as we are interested in the predictability of the upcoming playcall given numberous predictors. A logistic regression will be performed.

Table 1: Play Type Counts

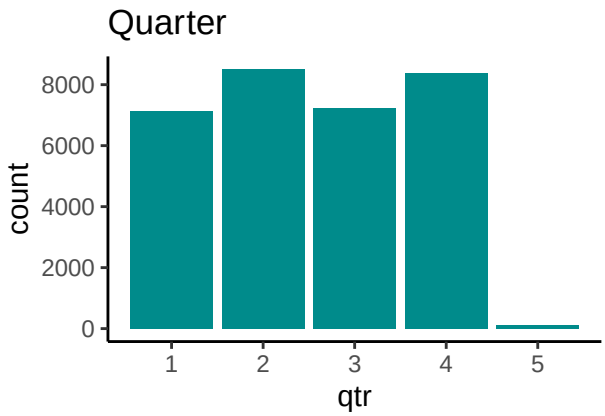
play_type	n	percentage
pass	19138	0.61001
run	12235	0.38999

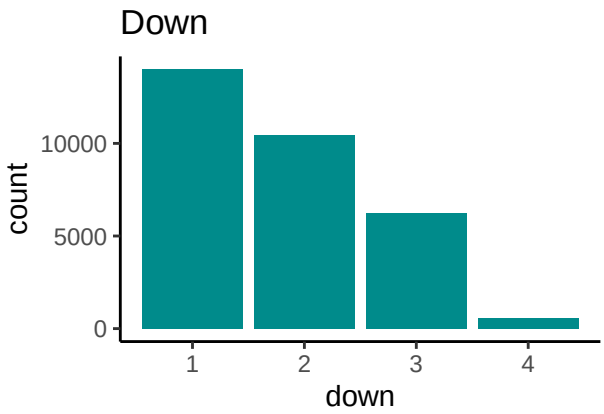


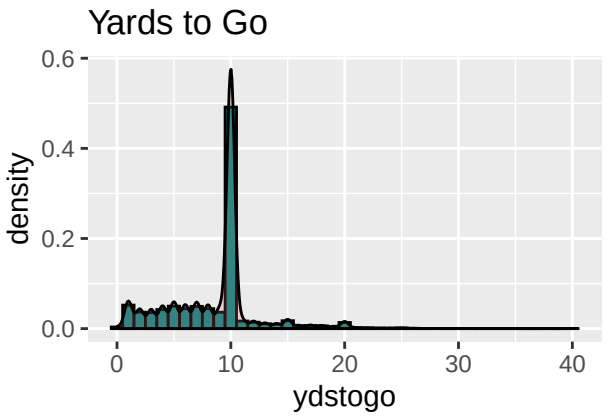
Analysis of Potential Predictors

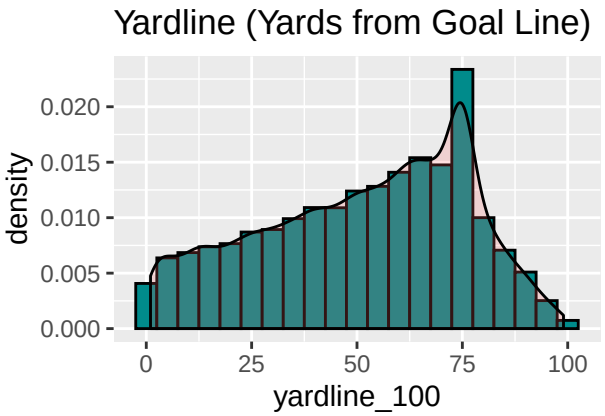
The next 6 figures show the distributions of the individual predictor variables. The first two charges are bar charts of the quarter (where a value of 5 represents overtime) and down. As expected, the distribution of plays across the 4 quarters are pretty evenly spread. Also expected is the downward trend as down increases, as in a cumulative sense, you must have more first downs than second downs, more second downs than thrid downs, and more third downs than fourth downs. The next two figures are the yards to go and yardline distributions.For yards to go, we see a spike at 10 because every NFL drive starts with a first and 10 yards to go. This is expected and the remaining values seem pretty evenly spread, outside of a few large outliers that may need to be excluded later on. The yardline distribution has a slightly skewed curve that is explained by majority of NFL drives starting on the 25 yard line (represented by a value of 75 in this variable). The last two charts are the Win Probability distribution and RB salary distribution. both of these are mostly symmetric and evenly spaced out, with some exceptions on the RB salary chart. With only 32 values this shape should be expected when binned by 0.5 (a true salary value of 500,000).

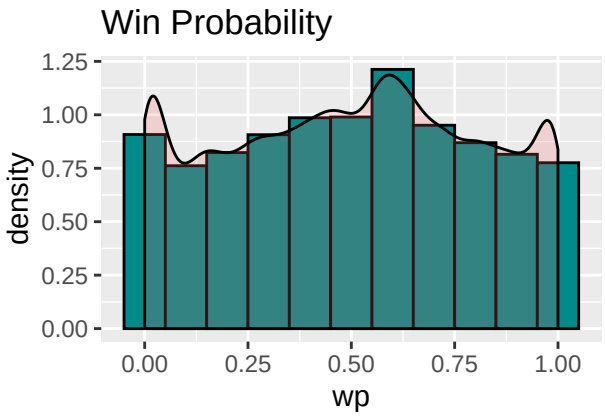
```
## Warning: Removed 110 rows containing non-finite values
## (stat_count).
```


Quarter
count
8000
6000
4000
2000
0
1
2
3
4
5
qtr


Down
count
10000
5000
0
1
2
3
4
down


Yards to Go
density
0.6
0.4
0.2
0.0
0
10
20
30
40
ydstogo


Yardline (Yards from Goal Line)
density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
yardline_100



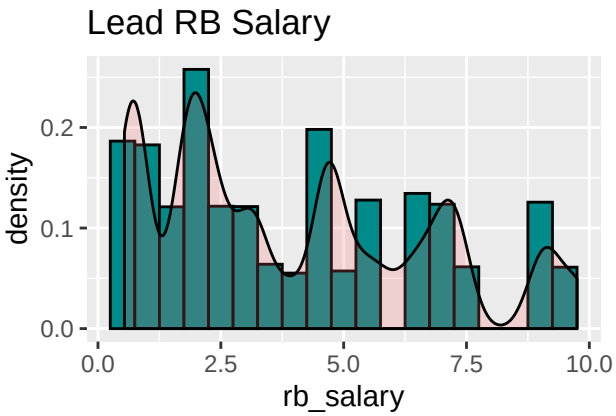


Table 2: Basic Statistics of Predictor Variables

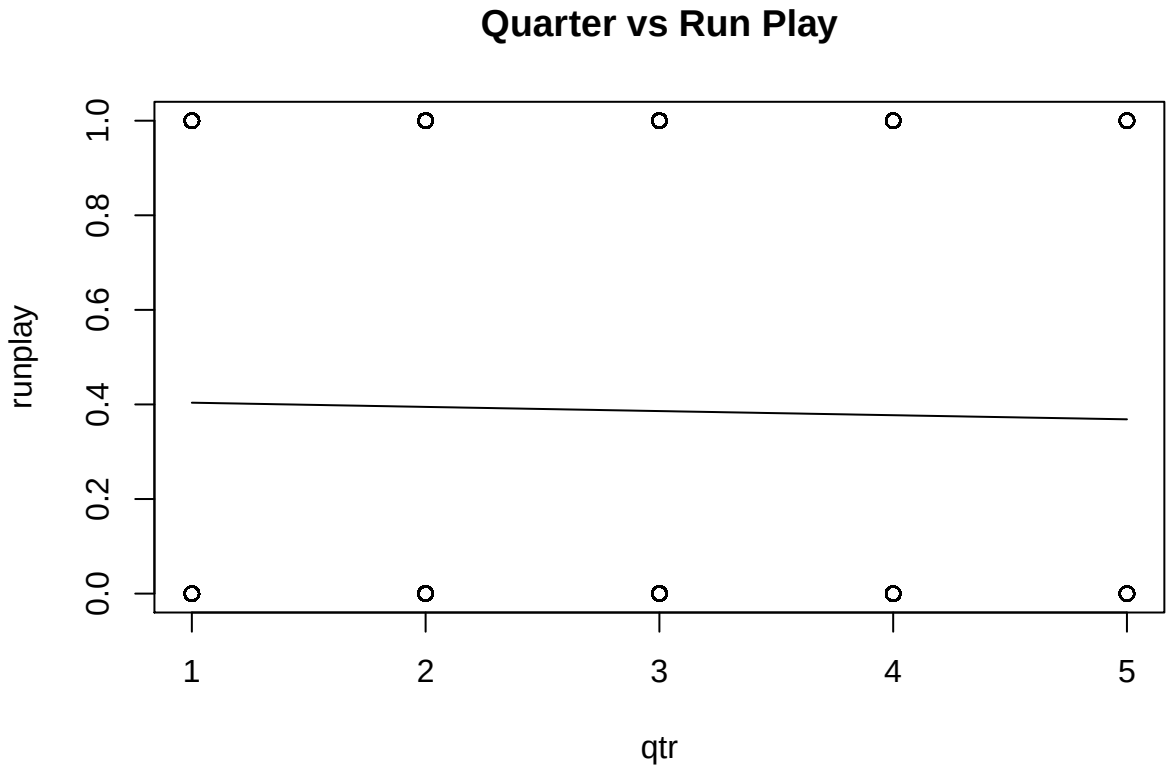
	qtr	down	ydstogo	yardline_100	wp	rb_salary
min	1.0000	1.00000	0.0000	1.000	0.00003	0.53814
max	5.0000	4.00000	40.0000	99.000	0.99993	9.75000
mean	2.5498	1.78854	8.6398	51.278	0.50010	3.97401
sd	1.1221	0.82231	4.0375	24.499	0.30078	2.70330
var	1.2592	0.67619	16.3013	600.222	0.09047	7.30783
mean	2.5498	1.78854	8.6398	51.278	0.50010	3.97401
IQR	2.0000	1.00000	4.0000	39.000	0.49426	4.48965

Effect of Quarter on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to data for the play type response variable and quarter. The entire dataset was used for the preliminary investigation. The quarter variable will be evaulated as both a numeric and as a factor, as the initial analysis shows that there is significance in quarters 1, 2 and 4 where a team may tend to call run plays less often. By itself, quarter may not be a strong predictor to differentiate when a team will run the ball.

Table 3: Coefficients

	x
(Intercept)	-0.29145
factor(qtr)2	-0.32588
factor(qtr)3	-0.03366
factor(qtr)4	-0.22720
factor(qtr)5	-0.28837



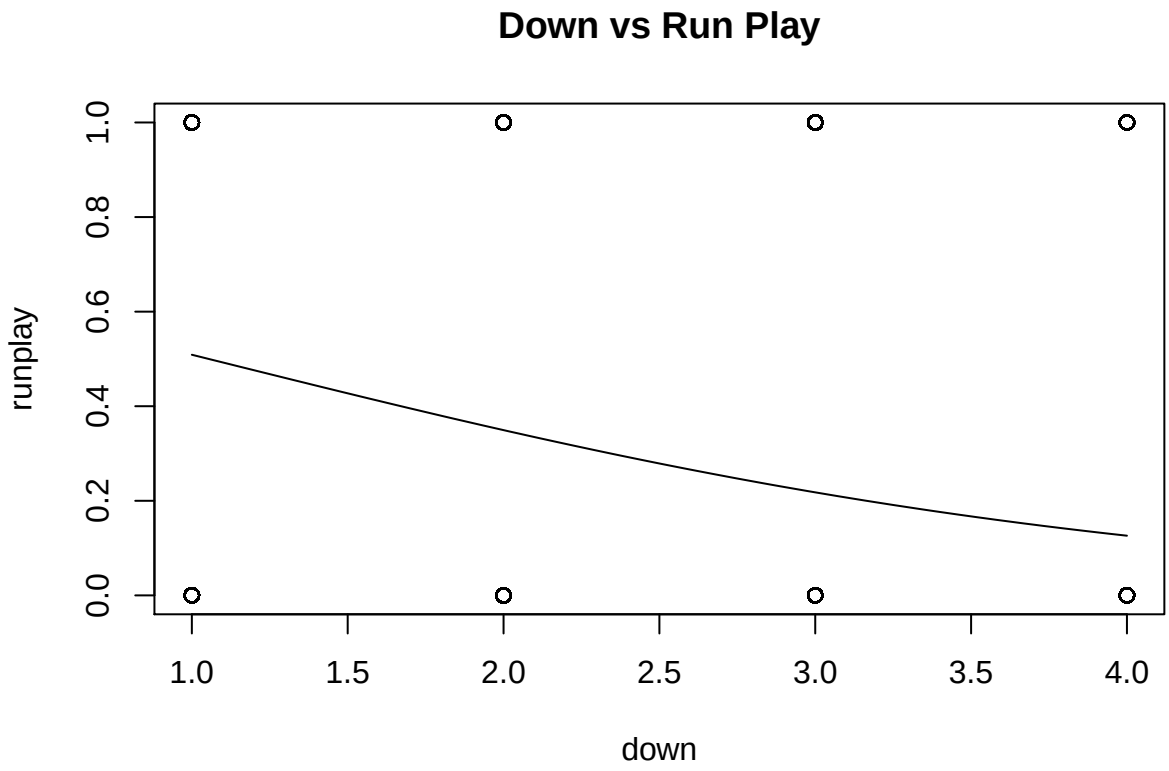

```
## Analysis of Deviance Table
##
## Model 1: runplay ~ qtr
## Model 2: runplay ~ factor(qtr)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     31371      41948
## 2     31368      41825  3      123   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effect of Down on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to the data for the play type reponse variable and down. In the plot we can see a negative trend as the down increases, meaning we can expect a pass play more often on later downs. Again, we will compare the down predictor as numeric and as a factor to cover our bases. As a factor we see that 3rd down has the largest coefficient of -1.604, meaning on 3rd down a team the odds of a running play are 0.201 times the odds of a passing play.

Table 4: Coefficients

	x
(Intercept)	0.01686
factor(down)2	-0.53606
factor(down)3	-1.60355
factor(down)4	-0.73084




```
## Analysis of Deviance Table
##
## Model 1: runplay ~ down
## Model 2: runplay ~ factor(down)
```

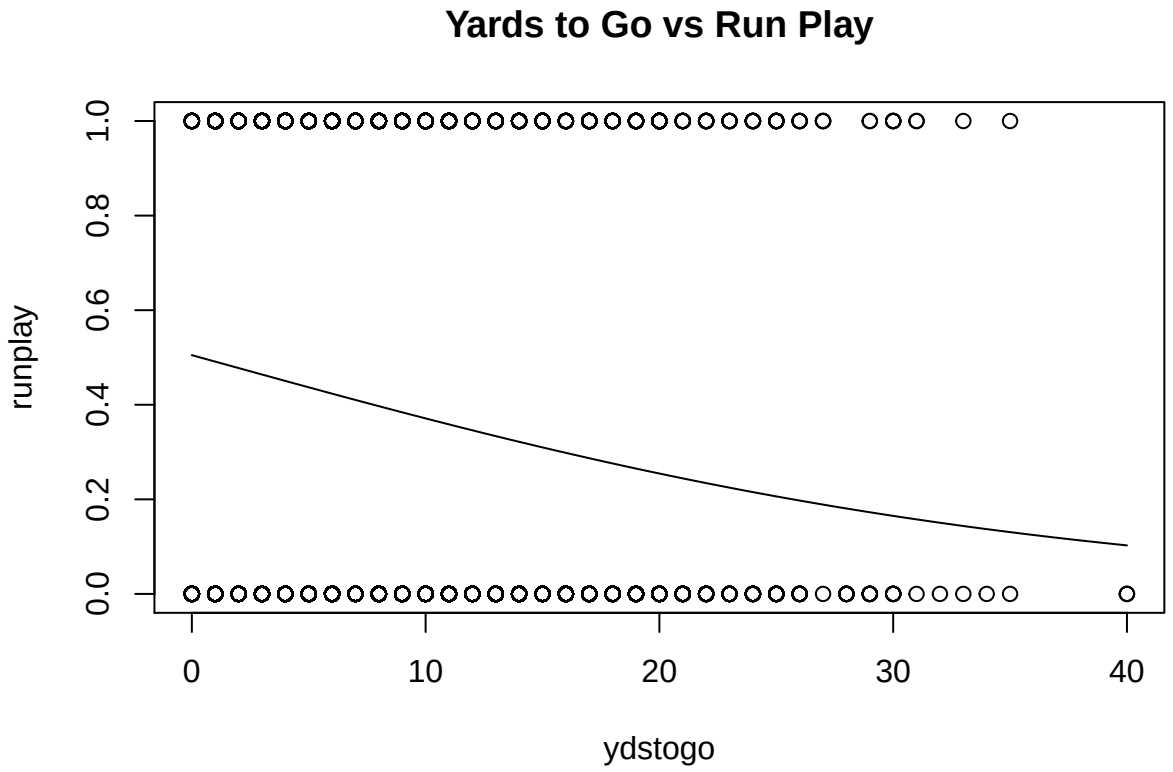
```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     31261      39904
## 2     31259      39629  2      275   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effect of Yards to Go on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to the data for the play type response variable and yards to go. Again we see a negative trend as the yards to go increases, the number of running plays called decreases. With a beta coefficient of -0.0547, for every yard increase the log odds ratio of 0.947 means that a team is 1.06 times more likely to call a pass play. This data may be skewed by a few outliers that do not represent an average NFL game. To improve the model, these data points greater than an arbitrary value may be removed.

Table 5: Coefficients

	x
(Intercept)	0.01967
ydstogo	-0.05470

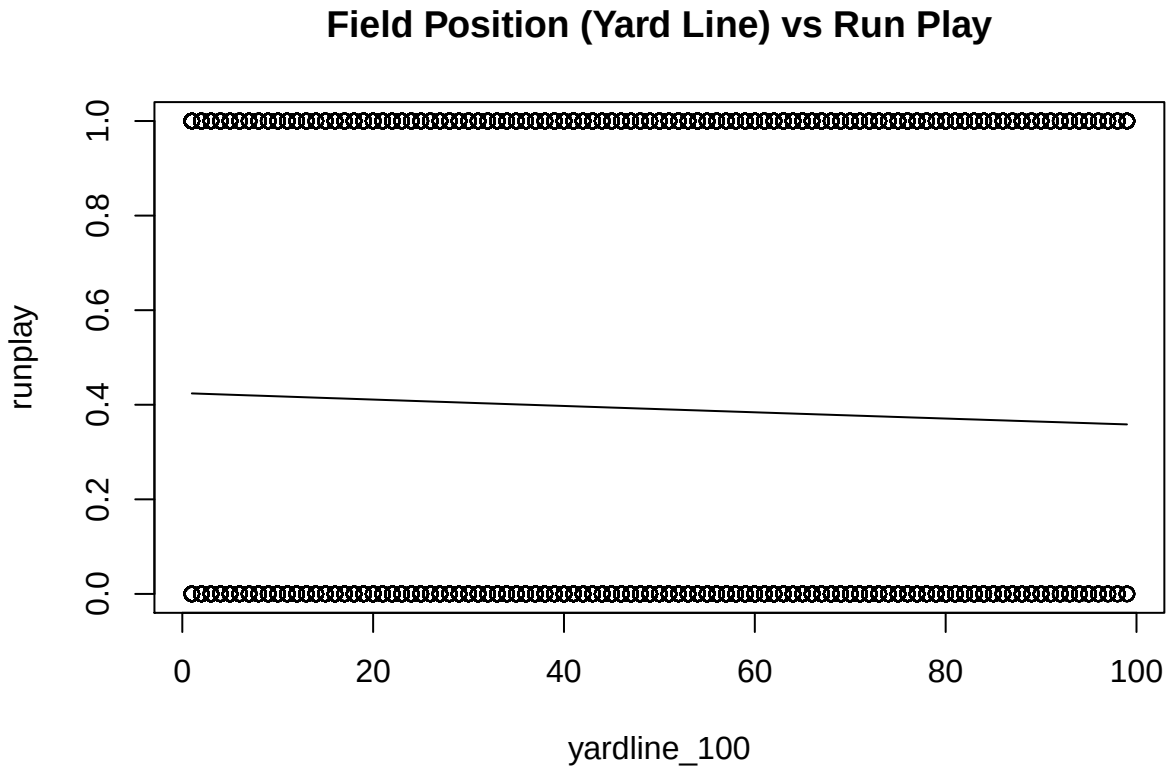


Effect of Yard Line on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to the data for the play type response variable and field position, represented by the yards away from the goal line. By itself, yard line has a slight negative relationship with the number of running plays, however it is not very strong as the beta coefficient was found to be -0.002813. With this being so close to zero, this variable by itself may not be useful.

Table 6: Coefficients

	x
(Intercept)	-0.30362
yardline_100	-0.00281

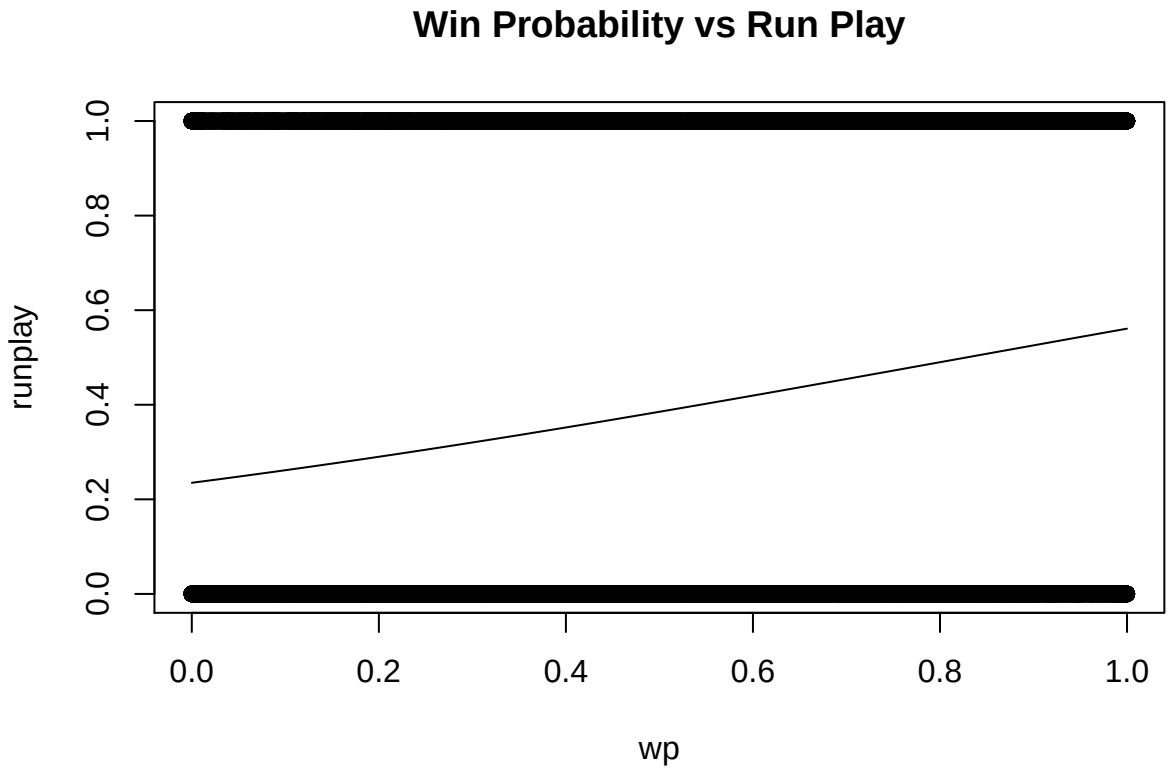


Effect of Win Probability on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to the data for the play type response variable and win probabiliy. This variable is the default NFLFastR win probability model that typically tracks with the Vegas model. There is a clear positive trend, and the beta coefficient of 1.425 gives a log odds ratio of 4.16. This lines up with as a team is winning by a lot or when a game is close to being over, a team will call a run play more often to run out the clock.

Table 7: Coefficients

	x
(Intercept)	-1.1804
wp	1.4253



Effect of RB Salary on Play Type

The below figure illustrates the scatter plot computed by the simple logistic regression model fit to the data for the play type response variable and RB salary. There is a very, very minor negative trend between these two variables. The beta coefficient of -0.015 produces a log odds ratio of 0.986, which is close to 1 which would suggest no real impact on the probabilty of a run play being called. Alone this predictor does not appear to be useful, but we will continue to evaluate in the final model.

Table 8: Coefficients

	x
(Intercept)	-0.38976
rb_salary	-0.01454

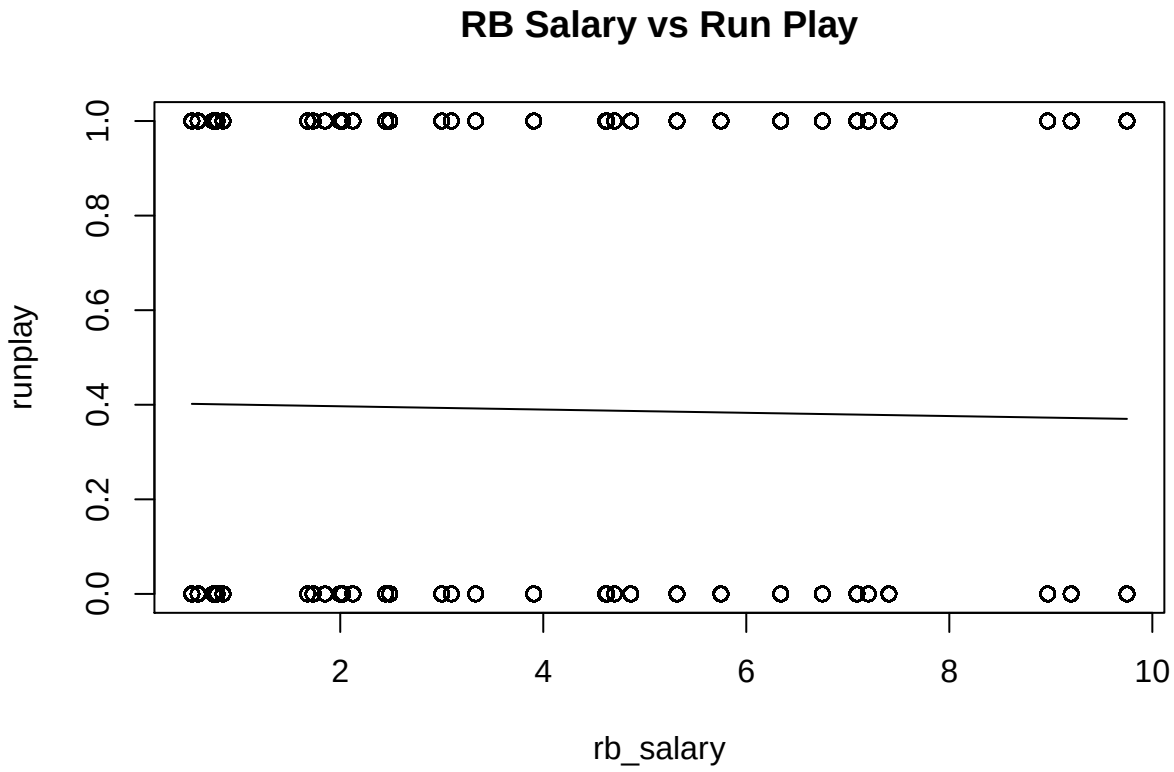


Table 9: Analysis of Inidividual Predictor Models

Predictor	Beta1	Log.Odds.Ratio	SE	Deviance	AIC
Quarter	-0.03710	0.964	0.01030	41825	41835
Down	-0.65680	0.519	0.01570	38629	39637
Yards to Go	-0.05470	0.947	0.00298	41612	41616
Yard Line	-0.00281	0.997	0.00047	41925	41929
WP	1.42530	4.160	0.04020	40650	40654
RB Salary	-0.01454	0.986	0.00429	41949	41953

Multicollinearity between Predictor variables :

The table below shows the correlation coefficiencts r between all of the variables considered for the final logistic regression model. There does not appear to be any mutlicollinearity issues in this dataset. No correlation was suspected, but as a thorough review this section is included.

Table 10: Correlation Coefficient

	qtr	down	ydstogo	yardline_100	wp	rb_salary
qtr	1.00000	NA	-0.00957	-0.04344	-0.07082	0.00430
down	NA	1	NA	NA	NA	NA
ydstogo	-0.00957	NA	1.00000	0.25551	-0.04474	0.00593
yardline_100	-0.04344	NA	0.25551	1.00000	-0.12601	0.00868
wp	-0.07082	NA	-0.04474	-0.12601	1.00000	-0.07899
rb_salary	0.00430	NA	0.00593	0.00868	-0.07899	1.00000

Model Selection

Automatic Variable Selection Method

For this project we will use a purposeful selection process proposed by Hosmer. In a shortened form, the steps are:

- Construct an initial main-effects model using predictors that show any evidence of being relevant
- Conduct backwards elimination, keeping a variable if it is either significant or shows evidence of being a relevant confounder
- Add additional variables that were not included in step 1 but are significant after step 2
- Check for plausible interactions among variables
- Conduct follow-up diagnostic investigations

The initial model consists of the Down and Win Probability predictors. These were chosen as their log odds ratio differed from 1 the most and their standard errors did not point towards questionable relevance.

Table 11: Purposeful Model Selection - All

Model	Predictors	Deviance	df	AIC	Compared	Dev.Diff
1	None	41961	31372	41963		
2	D + WP	38623	31260	38629	(2)-(1)	3338
3	D + WP + YTG	37187	31259	37195	(3)-(2)	1436
4	D + WP + YTG + fct(Q)	37060	31255	37076	(4)-(3)	127
5	D + WP + YTG + fct(Q) + YrdLine	37050	31254	37068	(5)-(4)	10
6	D + WP + YTG + fct(Q) + YrdLine + RBSal	37050	31253	37070	(6)-(5)	0
7	fct(D) + WP + YTG + fct(Q) + YrdLine	36757	31252	36779	(7)-(5)	293
8	fct(D) + YTG + fct(Q) + YrdLine + fct(D) * YTG	36632	31249	36660	(8)-(7)	125

Model	Predictors	Deviance	df	AIC	Compared	Dev.Diff
9	fct(D) + YTG + fct(Q) + YrdLine + fct(D) * YTG * YrdLine	36340	31242	36382	(9)-(8)	292

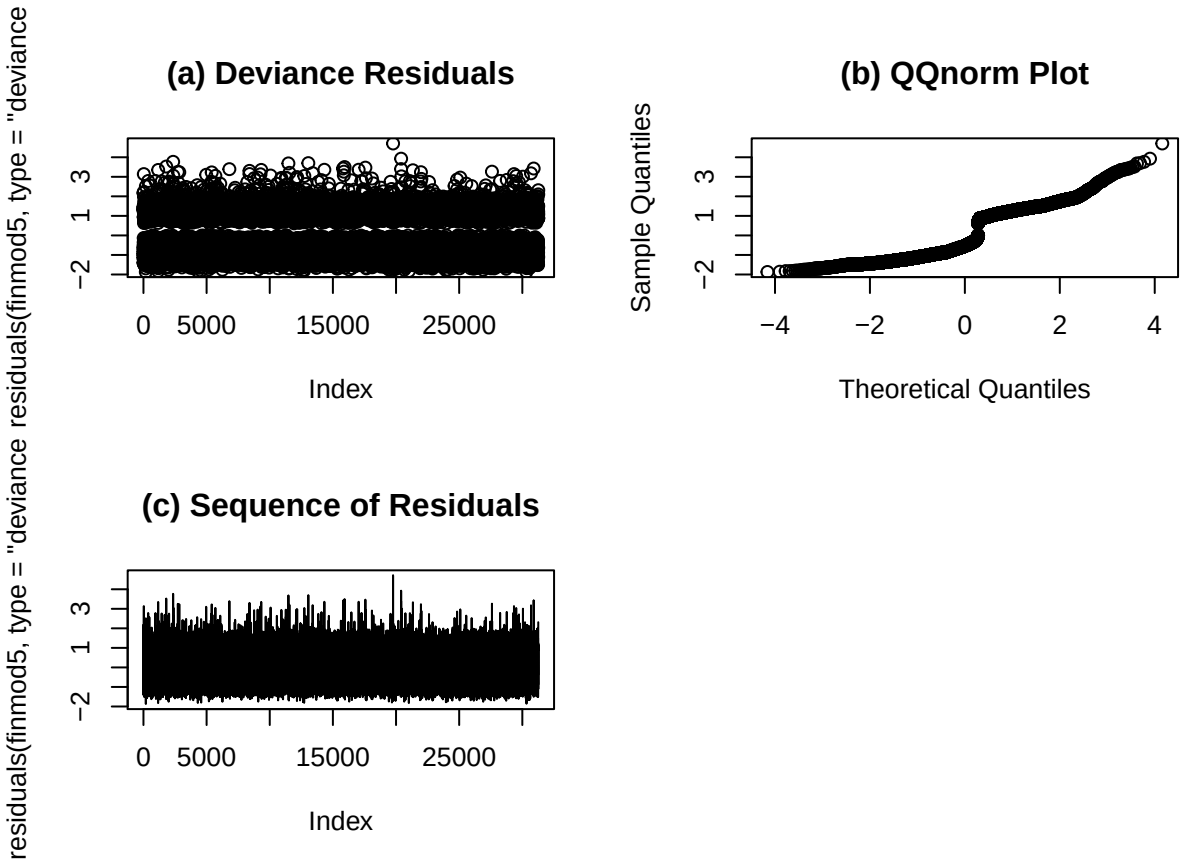
After evaluating a logistic model with all 6 potential predictors, a more simple approach using the Down, Yards to Go, Yard Line, and WP variables was completed. Any potential interaction terms with WP were omitted as interaction within this model already exists. This was done as the deviance and AIC values were extremely large, so a more simple model was decided upon. The beta values and their corresponding standard errors did not inspire confidence for a complex model containing 6 or more predictors. Below is an updated deviance chart depicting the final model selection.

Table 12: Purposeful Model Selection - Simple

Model	Predictors	Deviance	df	AIC	Compared	Dev.Diff
1	None	41961	31372	41963		
2	fct(D)	39629	31259	39637	(2)-(1)	2332
3	fct(D) + YTG	38152	31258	38162	(3)-(2)	1447
4	fct(D) + YTG + YrdLine	38152	31257	38164	(4)-(3)	0
5	fct(D) + YTG + fct(D) : YTG	38004	31255	38020	(5)-(3)	148
6	fct(D) + YTG + WP + fct(D) : YTG	36769	31254	36787	(6)-(5)	1235

Residual Diagnostics

The next three plots show a quick diagnostics of the resiudals. Figure (a) are the deviance residuals since standardized residuals are not to be used with ungrouped data. For this project, ungrouped data was used. Figure (b) is a plot evaluating normality. From these plots we do see some datapoints with large residuals and not perfect normality. The third plot helps show that the play calling did not change much as the season progressed.



Goodness of Fit Test

To fulfill the objective of this study, scatterplots, tables, logistic regression models, p-value, coefficient of determination, and likelihood ratio tests have been used.

All of the statistical analyses in this document will be performed using R version 3.6.1 (2019-07-05) R (R Core Team, 2018). R packages used will be maintained using the packrat dependency management system.

Results

For each of the variables in the dataset, we test if there exists an assocation with whether a run play was called using two-tailed Wald test statistic z individually. The z test considers the following hypothesis:

Null Hypothesis H_o : Beta_1 = 0

Alternative Hypothesis H_a : Beta_1 != 0

The decision is taken considering z* = b_1-Beta_1/SE(b_1)

where:

- z* is the test statistic for the z test

- b_1 is the observed coefficient
- Beta_1 is the expected coefficient of the fitted logistic regression model
- SE(b_1) is the sampling variability of b_1

The z* statistic is tested against z(1-alpha/2, df) where: - alpha is the level of significance = 0.05 - df is the degrees of freedom

If z* > z(1-alpha/2, df), H_0 is rejected else the decision is taken in the favor of H_a. The decision rule also considers the p-value. If p-value <= alpha the decision is to reject H_o else we fail to reject H_o.

We also review the Likelihood Ratio test for each individual predictor variable as well as the log odds ratio and compare it to a value of 1, which shows the level of impact that predictor variable has on the response variable.

Effect of Quarter on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis for the 1st, 3rd, and 4th quarter. With this variable we see steady negative coefficients that help show us play calling throughout the game. For the initial model selection process, quarter was included however it was ultimately not included in the final model in an effort to reduce complexity, variables, and steer away from the win probability model. We can see the log odds ratios broken down by quarter in the table below. The 2nd quarter appears the most pass-heavy, as the odds of a run play were 0.722 times the odds of a pass play. The 3rd quarter appears the most evenly split, as the odds of a run play were 0.967 times the odds of a pass play.

Table 13: Quarter as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	-0.29145	0.74718
factor(qtr)2	-0.32588	0.72189
factor(qtr)3	-0.03366	0.96690
factor(qtr)4	-0.22720	0.79676
factor(qtr)5	-0.28837	0.74948

Effect of Down on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis for every down except 1st down. With this variable we again see steady negative coefficients that help show us the play calling throughout an individual drive. For the entire model process, the down variable was included as it plays a major role in the individual play. We can see the log odds ratios below which tell us interesting notes such as the odds that a team runs on 3rd down is 0.201 times as likely as the odds that a team passes the ball. On first down we see the log odds ratio sit near 1, meaning the play calling on 1st down is the most mysterious.

Table 14: Down as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	0.01686	1.01701
factor(down)2	-0.53606	0.58505
factor(down)3	-1.60355	0.20118
factor(down)4	-0.73084	0.48150

Effect of Yards to Go on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis for the yards to go variable. The p-value and standard error also are reasonable to include this variable in both the initial and final model. We can see the log odds ratio below which shows a negative trend, translating to an inverse relationship between yards to go and odds of a run play.

Table 15: Yards to Go as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	0.01967	1.01987
ydstogo	-0.05470	0.94677

Effect of Yard Line on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis for the yards to go variable. While this variable appears significant to the model, this variable does not appear relevant to the model when you look at the coefficient and standard error. With a beta of -0.0028 and a log odds ratio of 0.997, this variable was not included in the initial nor final model.

Table 16: Field Position as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	-0.30362	0.73814
yardline_100	-0.00281	0.99719

Effect of Win Probability on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis. As an individual predictor, WP was the strongest in relation to the play type. An odds ratio over 4 was found, which means the odds of a run play occurring for a winning team are much higher than the odds of a winning team passing the ball. WP is an existing model where interaction terms are assumed to be in use, thus more

interaction terms in our model were not considered.

Table 17: Win Probability as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	-1.1804	0.30716
wp	1.4253	4.15931

Effect of RB Salary on Play Type

At significance level (alpha) = 0.05, the decision reached by the z-test was to fail to reject the null hypothesis for the yards to go variable, however RB salary was not included in the initial model as it's effect on play type appeared minimal at best. With a log odds ratio of 0.99, the decision was made to not use this as a predictor variable.

Table 18: Running Back Salary as a Sole Predictor

	Beta	Log.Odds.Ratio
(Intercept)	-0.38976	0.67722
rb_salary	-0.01454	0.98556

Primary Objective Results

The results from the individual z-tests and likelihood ratio tests show that an association exists between multiple of the potential predictor variables. However, the method of purposeful selection began with a simple model consiting of down as a factor. Then yards to go and win probability were added followed by their interaction terms. Each additional term reduced the deviance and AIC, which were the main values to compare models used in this study. At a 95% predictive interval, realistic log odds ratios were identified from the final model.

Discussion and Conclusion

The estimated regression function from this data analysis would be:

glm(formula = runplay ~ factor(down) * ydstogo * wp

Y-hat = 0.122 - 0.19c1 - 0.584c2 + 0.031c3 - 0.092x1 + 1.581x2 - 0.06c1x1 - 0.253c2x1 - 0.481c3x1 - 0.321c1x2 - 1.491c2x2 + 0.745c3x2 + 0.003x1x2 + 0.003c1x1x2 + 0.251c2x1x2 - 0.181c3x1x2

where, Y-hat is the play type, where 1 = run play c1 is the down factor, specifically 2nd down c2 is the down factor, specifically 3rd down c3 is the down factor, specifically 4th down X1 is the yards to go variable X2 is the WP variable c1x1 is the interaction term between 2nd down and yards to go c1x2 is the interaction term between 2nd down and win probability

The residual deviance for the final model described above is 36703, which is a fair decrease from the null model which had a residual deviance of 41827. This data proved challenging as the deviances for the simple logistic regressions proved to be very large and attempts at improving the model appeared to reduce the deviance in small chunks. This was somewhat anticipated, as each NFL team has an analytics department that reviews every aspect of the game to optimize their gameplan and get an edge over their opponent. If determining which play type was going to occur prior to it occurring was a simple task, that coach would not have a job for long. Another challenging part of this project was the limited variables made available to the public. Other potential predictors that have an influence are not readily available. Such predictors include defensive formation, offensive formation, individual coaching tendencies and much more. With the data made available from NFLFastR, this study shows that the predictor variables of down, yards to go, and win probability along with their respective interaction terms can predict whether an NFL team will run the ball. All statistical analysis is conducted at 95% confidence interval and at 0.05 significance level. All potential predictors in this data set were explored individually through the Wald z-test statistic and the Likelihood Ratio Test. Although only win probability showed significance and relevance, more parameters were included as domain knowledge directed the inclusion of them. Exploring the test statistics, it was found that the strongest association among variables is best explained by this final model. Win probability had the largest positive coefficient among predictors with a log odds ratio of 4.857. This tracks with the general sentiment that a team that is winning by a lot tends to run the ball more often to run the clock out and end the game sooner. This model also shows that on 3rd down the odds of a run play decrease significantly, with a log odds ratio of 0.558. Another impactful parameter in the final model is the interaction term between 3rd down and win probability. This parameter has a coefficient of -1.491 and a log odds ratio of 0.225. This aligns with the findings that 3rd down and win probability appear to be the two most impactful parameters in predicting whether a run call will be played.

Predictive Power

To look at the predictive power, we evaluate the final model at different downs and WP with 10 and 5 yards to go. 10 and 5 yards to go give us a general idea of the most common plays in an NFL game. As we can see, as down increases, WP increases, and yards to go increases, the odds of a run play decreases significantly.

Table 19: First Down and 10 - Differing WP

WP	Log.Odds.Ratio
0.25	0.67413
0.50	1.00776
0.75	1.50653

Table 20: Second Down and 10 - Differing WP

WP	Log.Odds.Ratio
0.25	0.28746
0.50	0.40431
0.75	0.56867

Table 21: Third Down and 10 - Differing WP

WP	Log.Odds.Ratio
0.25	0.03853
0.50	0.07434
0.75	0.14341

Table 22: Fourth Down and 10 - Differing WP

WP	Log.Odds.Ratio
0.25	0.00436
0.50	0.00500
0.75	0.00574

Table 23: First Down and 5 - Differing WP

WP	Log.Odds.Ratio
0.25	1.0635
0.50	1.5844
0.75	2.3603

Table 24: Second Down and 5 - Differing WP

WP	Log.Odds.Ratio
0.25	0.60680
0.50	0.84236
0.75	1.16935

Table 25: Third Down and 5 - Differing WP

WP	Log.Odds.Ratio
0.25	0.15759
0.50	0.22133
0.75	0.31087

Table 26: Fourth Down and 5 - Differing WP

WP	Log.Odds.Ratio
0.25	0.09535
0.50	0.13658
0.75	0.19563

Personal Takeaways and Notes

Initially I was disappointed with the findings from this study due to the large values in deviance, deviance residuals and a seemingly poor fit of my final model. After looking at the data more, I did have the following personal takeaways that were interesting:

- Running Back salary appeared to have little or no effect on whether a run play would be called. An initial assumption I had was that a team that spent more money on a running back would call more running plays. However, run plays appear to be mostly based on game script.
- The 2nd quarter appeared the most pass heavy, with a log odds ratio of 0.722. The 3rd quarter appeared the most run heavy, with a log odds ratio of 0.967.
- 1st down appeared the most evenly split, with a log odds ratio of 1.02. This makes sense as there is less risk in an unsuccessful play on first down in comparison to the successive downs. 3rd down is the most pass heavy, while 4th down also shows a favor towards passing the ball, just not as frequent. This is surmised to be from many of the 4th down plays being short distance, in which a running play could succeed.
- Surprisingly, field position did not impact the play calling significantly. With a log odds ratio of 0.998, this is not much of a difference maker. If this study was repeated, I think grouping the field position would yield better results. This way we could evaluate redzone play calling vs the rest of the field.

Appendix: R-code

```
# libraries used
library(knitr)
library(readxl)
```

```
library(formatR)
library(caret)
library(Hmisc)
library(stargazer)
library(xtable)
library(leaps)
library(ggplot2)
library(gridExtra)

# load NFL play by play data
pbp_2019 <- read.csv("C:/Users/Brenton/Desktop/KUMC/STAT835/Final/pbp_2019.csv",
    header = TRUE)
runpass <- filter(pbp_2019, qb_scramble == 0 & play_type == "run" |
    play_type == "pass")
nfldata <- select(runpass, play_type, yardline_100, qtr, down,
    ydstogo, wp, rb_salary, game_id, posteam, defteam, desc,
    goal_to_go)
nfldata$ydstogo <- ifelse(nfldata$goal_to_go == 0, nfldata$ydstogo,
    nfldata$yardline_100)
nfldata$down <- as.numeric(nfldata$down)
nfldata$qtr <- as.numeric(nfldata$qtr)
nfldata$runplay <- ifelse(nfldata$play_type == "run", 1, 0)

d1 <- nfldata %>% group_by(play_type) %>% tally()
d1$percentage <- d1$n/sum(d1$n)
d1
# Visualize response variable, runplay
p <- ggplot(d1, aes(x = play_type, y = n, fill = play_type)) +
    geom_bar(stat = "identity") + scale_fill_brewer(palette = "Set1") +
    theme(legend.position = "none") + ggtitle("Play Calling for 2019")

# evaluate predictor variables
qtrplot <- ggplot(nfldata, aes(qtr)) + geom_bar(fill = "darkcyan") +
    theme_classic() + ggtitle("Quarter")

downplot <- ggplot(nfldata, aes(down)) + geom_bar(fill = "darkcyan") +
    theme_classic() + ggtitle("Down")

ydstogoplot <- ggplot(nfldata, aes(x = ydstogo)) + geom_histogram(aes(y = ..density..),
    binwidth = 1, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
    fill = "#FF6666") + ggtitle("Yards to Go")

yardlineplot <- ggplot(nfldata, aes(x = yardline_100)) + geom_histogram(aes(y = ..densi
    binwidth = 5, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
```

```
    fill = "#FF6666") + ggtitle("Yardline (Yards from Goal Line)")

wpplot <- ggplot(nfldata, aes(x = wp)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.1, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
    fill = "#FF6666") + ggtitle("Win Probability")

rbsalplot <- ggplot(nfldata, aes(x = rb_salary)) + geom_histogram(aes(y = ..density..),
    binwidth = 0.5, color = "black", fill = "darkcyan") + geom_density(alpha = 0.2,
    fill = "#FF6666") + ggtitle("Lead RB Salary")

grid.arrange(qtrplot, downplot, ncol = 2, nrow = 2)
grid.arrange(ydstogoplot, yardlineplot, ncol = 2, nrow = 2)
grid.arrange(wpplot, rbsalplot, ncol = 2, nrow = 2)

nfldata2 <- select(nfldata, qtr, down, ydstogo, yardline_100,
    wp, rb_salary)
tabletry1 <- sapply(nfldata2, each(min, max, mean, sd, var, mean,
    IQR), na.rm = TRUE)
kable(tabletry1, caption = "Basic Statistics of Predictor Variables")


# runplay ~ qtr
qtrmod <- glm(runplay ~ qtr, family = binomial, data = nfldata)
qtrmod2 <- glm(runplay ~ factor(qtr), family = binomial, data = nfldata)
plot(runplay ~ qtr, nfldata, main = "Quarter vs Run Play")
curve(predict(qtrmod, data.frame(qtr = x), type = "resp"), add = TRUE)

anova(qtrmod, qtrmod2, test = "LRT")

# runplay ~ down
downmod <- glm(runplay ~ down, family = binomial, data = nfldata)
downmod2 <- glm(runplay ~ factor(down), family = binomial, data = nfldata)
plot(runplay ~ down, nfldata, main = "Down vs Run Play")
curve(predict(downmod, data.frame(down = x), type = "resp"),
    add = TRUE)

anova(downmod, downmod2, test = "LRT")

# runplay ~ yards to go
ytgmod <- glm(runplay ~ ydstogo, family = binomial, data = nfldata)
plot(runplay ~ ydstogo, nfldata, main = "Yards to Go vs Run Play")
curve(predict(ytgmod, data.frame(ydstogo = x), type = "resp"),
    add = TRUE)
```

```
# runplay ~ yard line
ydlinemod <- glm(runplay ~ yardline_100, family = binomial, data = nfldata)
plot(runplay ~ yardline_100, nfldata, main = "Field Position (Yard Line) vs Run Play")
curve(predict(ydlinemod, data.frame(yardline_100 = x), type = "resp"),
    add = TRUE)

# runplay ~ win probability
wpmod <- glm(runplay ~ wp, family = binomial, data = nfldata)
plot(runplay ~ wp, nfldata, main = "Win Probability vs Run Play")
curve(predict(wpmod, data.frame(wp = x), type = "resp"), add = TRUE)

# runplay ~ rb salary
rbsmod <- glm(runplay ~ rb_salary, family = binomial, data = nfldata)
plot(runplay ~ rb_salary, nfldata, main = "RB Salary vs Run Play")
curve(predict(rbsmod, data.frame(rb_salary = x), type = "resp"),
    add = TRUE)

# Table for predictors
modelframe <- data.frame(Predictor = c("Quarter", "Down", "Yards to Go",
    "Yard Line", "WP", "RB Salary"), Beta1 = c(-0.0371, -0.6568,
    -0.0547, -0.002813, 1.4253, -0.01454), `Log Odds Ratio` = c(0.964,
    0.519, 0.947, 0.997, 4.16, 0.986), SE = c(0.0103, 0.0157,
    0.00298, 0.000472, 0.0402, 0.00429), Deviance = c(41825,
    38629, 41612, 41925, 40650, 41949), AIC = c(41835, 39637,
    41616, 41929, 40654, 41953))
kable(modelframe, caption = "Analysis of Inidividual Predictor Models")

# Multicollinearity check
nfldata2 <- select(nfldata, qtr, down, ydstogo, yardline_100,
    wp, rb_salary)
kable(cor(nfldata2), caption = "Correlation Coefficient")

# Complex Model Selection
nullmod <- glm(runplay ~ 1, family = binomial, nfldata)
downmod <- glm(runplay ~ down, family = binomial, nfldata)
model2 <- glm(runplay ~ down + wp, family = binomial, nfldata)
model3 <- glm(runplay ~ down + wp + ydstogo, family = binomial,
    nfldata)
model4 <- glm(runplay ~ down + wp + ydstogo + factor(qtr), family = binomial,
    nfldata)
model5 <- glm(runplay ~ down + wp + ydstogo + factor(qtr) + yardline_100,
    family = binomial, nfldata)
model6 <- glm(runplay ~ down + wp + ydstogo + factor(qtr) + yardline_100 +
    rb_salary, family = binomial, nfldata)
```

```
model7 <- glm(runplay ~ factor(down) + wp + ydstogo + factor(qtr) +
    yardline_100, family = binomial, nfldata)
model8 <- glm(runplay ~ factor(down) * ydstogo + factor(qtr) +
    yardline_100 + wp, family = binomial, nfldata)
model9 <- glm(runplay ~ factor(down) * ydstogo * yardline_100 +
    factor(qtr) + wp, family = binomial, nfldata)

# Simple Model Selection
finmod1 <- glm(runplay ~ factor(down), family = binomial, nfldata)
finmod2 <- glm(runplay ~ factor(down) + ydstogo, family = binomial,
    nfldata)
finmod3 <- glm(runplay ~ factor(down) + ydstogo + yardline_100,
    family = binomial, nfldata)
finmod4 <- glm(runplay ~ factor(down) * ydstogo, family = binomial,
    nfldata)

summary(models)

simpleselect <- data.frame(Model = c(1, 2, 3, 4, 5), Predictors = c("None",
    "fct(D)", "fct(D) + YTG", "fct(D) + YTG + YTG", "fct(D) + YTG + fct(D)*YTG"),
    Deviance = c(41961, 39629, 38152, 38152, 38004), df = c(31372,
        31259, 31258, 31257, 31255), AIC = c(41963, 39637, 38162,
        38164, 38020), `Models Compared` = c(" ", "(2)-(1)",
        "(3)-(2)", "(4)-(3)", "(5)-(3)"), `Deviance Difference` = c(" ",
        2332, 1447, 0, 148))
kable(purposelect, caption = "Purposeful Model Selection - Simple")

# Residual Diagnostics
par(mfrow = c(2, 2))
plot(rstandard(finmod4), main = "(a) Standardized Residuals")
qqnorm(residuals(finmod4), main = "(b) QQnorm Plot")
plot(residuals(finmod4), type = "l", main = "(c) Sequence of Residuals")

# Results - Predictors
qtreval <- data.frame(Beta = qtrmod2$coefficients, `Log Odds Ratio` = exp(qtrmod2$coeffi
kable(qtreval, caption = "Quarter as a Sole Predictor")
downeval <- data.frame(Beta = downmod2$coefficients, `Log Odds Ratio` = exp(downmod2$coe
kable(downeval, caption = "Down as a Sole Predictor")
ytgeval <- data.frame(Beta = ytgmod$coefficients, `Log Odds Ratio` = exp(ytgmod$coeffici
kable(ytgeval, caption = "Yards to Go as a Sole Predictor")
yrdlneval <- data.frame(Beta = ydlinemod$coefficients, `Log Odds Ratio` = exp(ydlinemod$
kable(yrdlneval, caption = "Field Position as a Sole Predictor")
wpeval <- data.frame(Beta = wpmod$coefficients, `Log Odds Ratio` = exp(wpmod$coefficient
kable(wpeval, caption = "Win Probability as a Sole Predictor")
```

```
rbseval <- data.frame(Beta = rbsmod$coefficients, `Log Odds Ratio` = exp(rbsmod$coeffici
kable(rbseval, caption = "Running Back Salary as a Sole Predictor")

# Response Variable Prediction
fmf <- data.frame(Predicted = c(mean(fitted(finmod6))), Observed = c(mean(nfldata$runpla
kable(fmf, caption = "Predicted vs Observed")

# Predictive Power -- FIX
c1 <- c(exp(predict(finmod6, data.frame(down = 1, ydstogo = 10,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 1, ydstogo = 10, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 1,
    ydstogo = 10, wp = 0.75), level = 0.95, interval = "prediction")))

fdf <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c1)

kable(fdf, caption = "First Down and 10 - Differing WP")

c2 <- c(exp(predict(finmod6, data.frame(down = 2, ydstogo = 10,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 2, ydstogo = 10, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 2,
    ydstogo = 10, wp = 0.75), level = 0.95, interval = "prediction")))

fdf2 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c2)

kable(fdf2, caption = "Second Down and 10 - Differing WP")

c3 <- c(exp(predict(finmod6, data.frame(down = 3, ydstogo = 10,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 3, ydstogo = 10, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 3,
    ydstogo = 10, wp = 0.75), level = 0.95, interval = "prediction")))

fdf3 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c3)

kable(fdf3, caption = "Third Down and 10 - Differing WP")

c4 <- c(exp(predict(finmod6, data.frame(down = 4, ydstogo = 10,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 4, ydstogo = 10, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 4,
    ydstogo = 10, wp = 0.75), level = 0.95, interval = "prediction")))
```

```
fdf4 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c4)

kable(fdf4, caption = "Fourth Down and 10 - Differing WP")

c51 <- c(exp(predict(finmod6, data.frame(down = 1, ydstogo = 5,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 1, ydstogo = 5, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 1,
    ydstogo = 5, wp = 0.75), level = 0.95, interval = "prediction")))

fdf51 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c51)

kable(fdf51, caption = "First Down and 5 - Differing WP")

c52 <- c(exp(predict(finmod6, data.frame(down = 2, ydstogo = 5,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 2, ydstogo = 5, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 2,
    ydstogo = 5, wp = 0.75), level = 0.95, interval = "prediction")))

fdf52 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c52)

kable(fdf52, caption = "Second Down and 5 - Differing WP")

c53 <- c(exp(predict(finmod6, data.frame(down = 3, ydstogo = 5,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 3, ydstogo = 5, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 3,
    ydstogo = 5, wp = 0.75), level = 0.95, interval = "prediction")))

fdf53 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c53)

kable(fdf53, caption = "Third Down and 5 - Differing WP")

c54 <- c(exp(predict(finmod6, data.frame(down = 4, ydstogo = 5,
    wp = 0.25), level = 0.95, interval = "prediction")), exp(predict(finmod6,
    data.frame(down = 4, ydstogo = 5, wp = 0.5), level = 0.95,
    interval = "prediction")), exp(predict(finmod6, data.frame(down = 4,
    ydstogo = 5, wp = 0.75), level = 0.95, interval = "prediction")))

fdf54 <- data.frame(WP = c(0.25, 0.5, 0.75), `Log Odds Ratio` = c54)

kable(fdf54, caption = "Fourth Down and 5 - Differing WP")
```

References

R Core Team (2018), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Dean, A., Voss, D., and Draguljic, D. (2017), Design and Analysis of Experiments, 2nd ed., Springer.

Kuehl, R. O. (2000), Design of Experiments: Statistical Principles of Research Design and Analysis, 2nd ed., Brooks/Cole, Cengage Learning.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin.

https://github.com/guga31bb/nflfastR-data , NFLFastR team

Bibliography

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.