Analyzing NFL Red Zone Efficiency: Discovering Strategies for Optimal Scoring Output

Timothy French, Umer Amir

2024-05-10

1.0 Introduction

1.1 Project Description The main guiding question for this project is: "How can NFL teams optimize their offensive output in the red zone?" The input variables in our study were the type of play ran, meaning either pass or run, and play formation. These variables were associated with the outcomes of yards gained, as well as if the play resulted in a first down or a touchdown. Therefore, some smaller sub-questions to guide our project are: Are pass or run plays more likely to gain yards, score a touchdown, or achieve a first down in the red zone? What formations are best for achieving these outputs in the red zone as well?

The domain of this research is professional sports, and more specifically, American Football in the United States (NFL). The data gathered is play-by-play data from every NFL game over the past five seasons (2019-2023). These are important questions to answer in terms of the success of an NFL team. An offense can be efficient in gaining yards and getting to the red zone, but if they cannot score once there, their drive stalls out and there chance of winning decreases. The difference between scoring a touchdown in the red zone and having to settle for a field goal can be the difference between winning and losing many games in an NFL season.

1.2 Background and Essential Terms:

- Yards: An offense's goal in football is typically to gain enough yards to reach the opposing team's end zone (defined below) in any way possible
- Pass vs. Rush vs. Scramble plays: In a pass play, a player, usually the quarterback, passes the ball forward in an attempt to gain yards. In a rush play, there is no forward pass. The quarterback can hand the ball off to another player, or can run on a play that was designed as a pass, known as a scramble, meaning a scramble is sub-category of a rush play.
- Formation: In football, the way an offense positions their players before the play begins is known as the formation. The various formations at their base levels are explained in section 2.1.
- End Zone, Touchdown: If a player reaches the opposing teams end zone with the ball, a touchdown is scored (worth 6 points)
- Red Zone: The part of the field from inside the opponent's 20 yard line to their goal line.
- Series First Down: 10 yards are needed for an offense's set of four downs to reset. If an offense gains ten yards before those four downs are over, they gain a first down and the downs are reset.
- Field Goal: the offensive team attempts to kick the ball between the field goal posts to gain three points if made

2.0 Data Description

Sections 2.1: Data Fields

The SeriesFirstDown, Formation, PlayType, IsRush, IsPass, and IsTouchdown were all manually converted into type Factor. Their levels will be displayed below.

SeriesFirstDown - play resulted in first down or not (1 = first down, 0 = no first down)
Yards - yards gained or lost (numeric)
Formation - formation offense begins play in - Value 1: "SHOTGUN" (qb lined up a short distance behind the center)
- Value 2: "UNDER CENTER" (qb directly behind the center)
- Value 3: "NO HUDDLE" (team does not huddle up before getting into formation)
- Value 4: NO HUDDLE SHOTGUN
- Value 5: WILDCAT (non-qb shotgun)
PlayType - type of play ran by offense
- Value 1: "PASS"
- Value 2: "RUSH"
- Value 3: "SCRAMBLE"
IsRush - offense ran rush play or not (1 = rush, 0 = not a rush)
IsPass - offense ran pass play or not (1 = pass, 0 = not a pass)
IsTouchdown - play resulted in offensive touchdown (1 = touchdown, 0 = not a touchdown)
YardLineFixed - numeric: yard line the play starts on
YardLineDirection - play ran in own territory (before 50 yard line) or opponent territory (after 50 yard line)
- Value 1: "OPP" (opponent)
- Value 2: "OWN"

Originally, we were planning on using a variable that took into account pass direction/distance and a variable that took into account rush direction. However, each of these variables had so many possible outcomes and that they made the outputs of our models difficult to interpret. They would be useful to NFL teams, but should be studied individually in another study for better readability and interpretability.

Other outcome variables in the original data set such as IsPenalty, IsIncomplete, IsSack, IsInterception, and IsFumble, could have been considered. However, they were excluded as this study generally focuses on which input variables mentioned result in positive outcomes in the red zone, where as each of these variables are associated with negative outcomes. In addition, they were excluded in order to simplify the scope of the study.

The original data contained five data sets:
2019 Play-by-Play Data Set: 42186 obs. of 45 variables
2020 Play-by-Play Data Set: 46189 obs. of 45 variables
2021 Play-by-Play Data Set: 42795 obs. of 45 variables
2022 Play-by-Play Data Set: 38598 obs. of 45 variables
2023 Play-by-Play Data Set: 39472 obs. of 45 variables

These five seasons of data were combined into one data set, resulting in a data frame containing 205,237 observations of 45 variables. Then, we excluded any observations of the data frame that were not plays ran in the red zone. To achieve this, we excluded any plays (observations) in which the YardLineFixed variable was 20 or greater, and in which the YardLineDirection variable was not "OPP". This made it so our data frame only included plays ran in opponent territory with 19 or less yards between the yard line in which the play was ran and the end zone. The data frame still needed to be cleaned as some of the plays included were still not red zone plays. To resolve this, we excluded any plays were not either a rush or a pass (meaning a play was ran), and plays in which more than 19 yards were gained (meaning the play did not start in the red zone). At the end of this process, our data frame to work with contained 20,773 observations.

2.2: Distributions of Variables Important for Analysis

IsTouchdown

Table 1: Touchdown Frequency

Touchdown	Count
No Touchdown	16461
Touchdown	4312

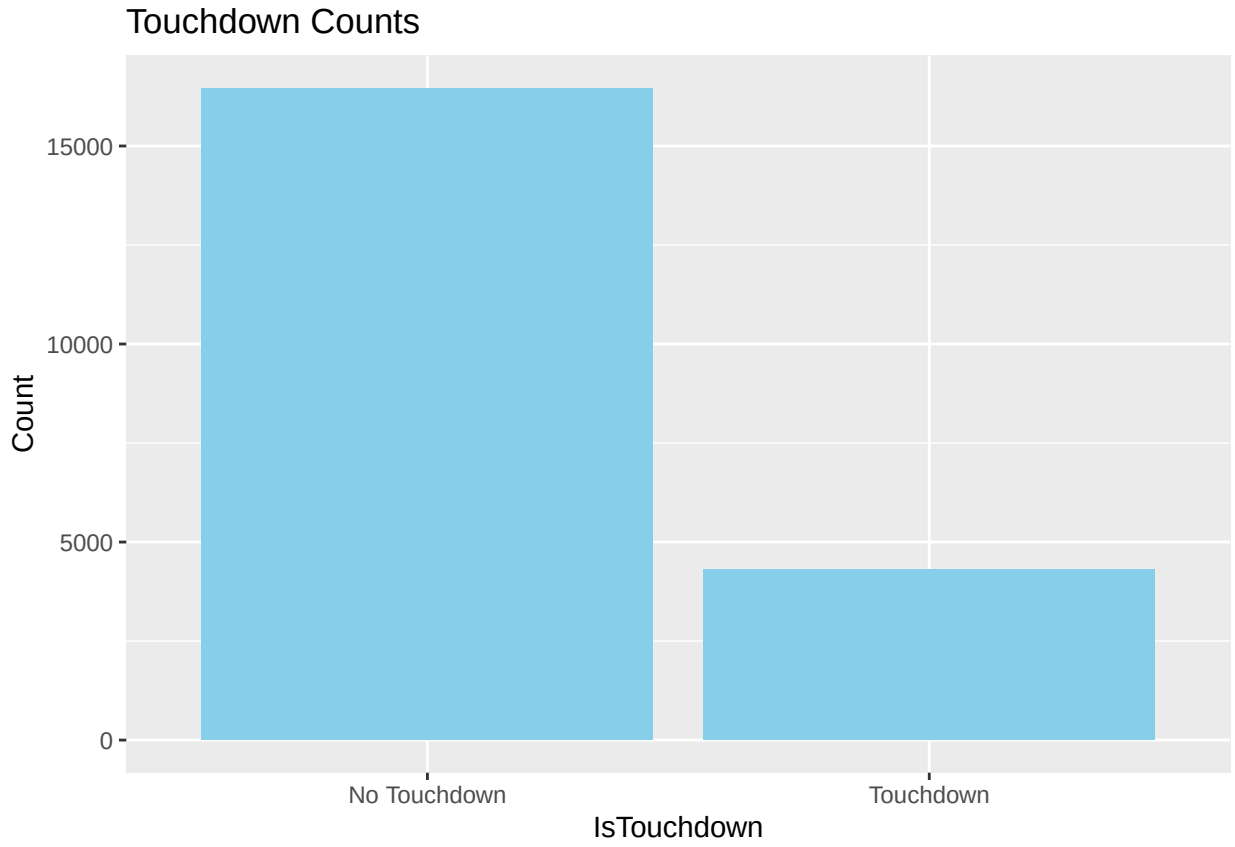

Figure 1: Touchdown Frequency Plot

Mild Imbalance: Touchdown = minority class (20.8%)
As would be expected, a minority of plays ran in the red zone actually result in touchdowns (4312/20773).

SeriesFirstDown

Table 2: First Down Frequency

First Down	Count
No First Down	14521
First Down	6252

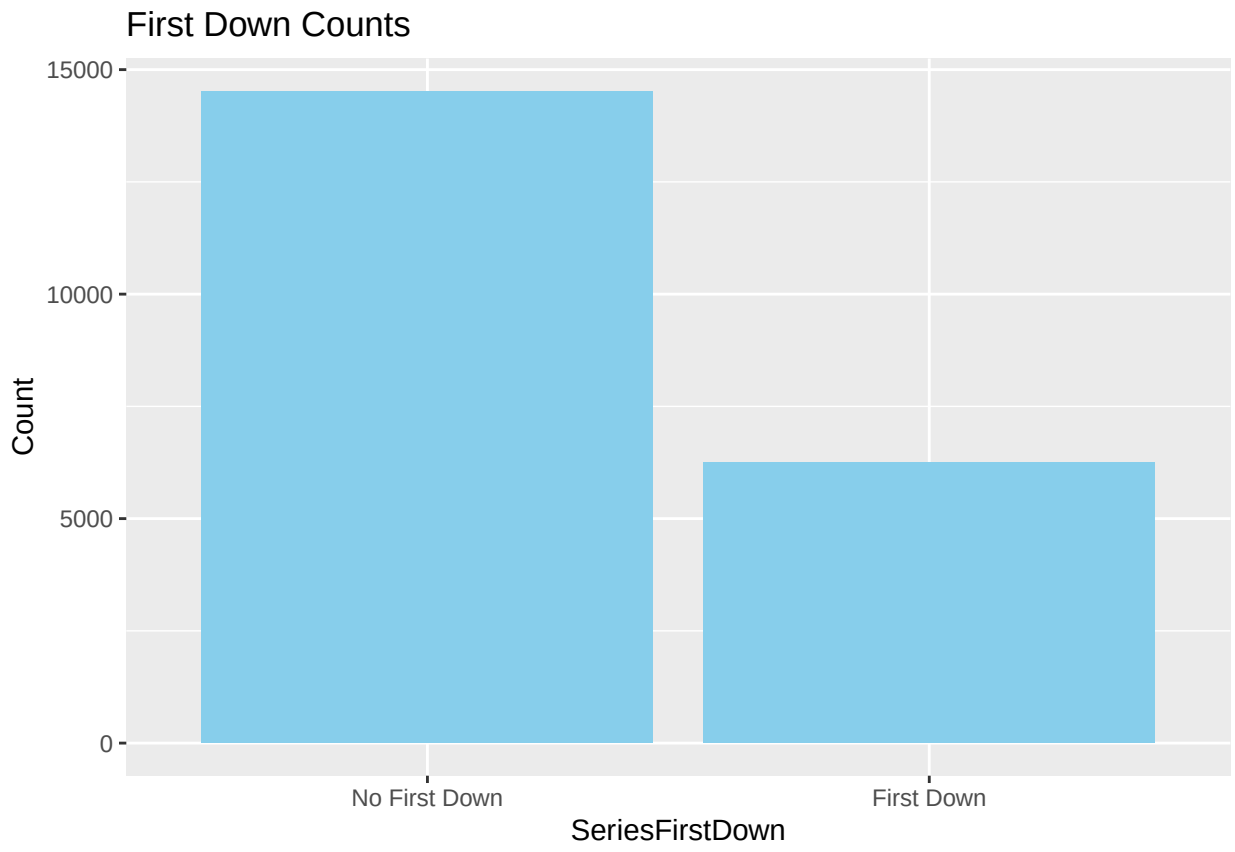

Figure 2: First Down Frequency Plot

Mild Imbalance: First Down = minority class (30.1%)
As would be expected, a minority of plays ran in the red zone actually result in first downs (6252/20773). In addition, it makes sense that more plays result in first downs rather than touchdowns as a defense's number one goal is to protect the end zone and prevent the offense from scoring.

IsRush and IsPass

Table 3: Rush and Pass Frequencies

Rush/Pass	Count
Pass	10964
Rush	9809

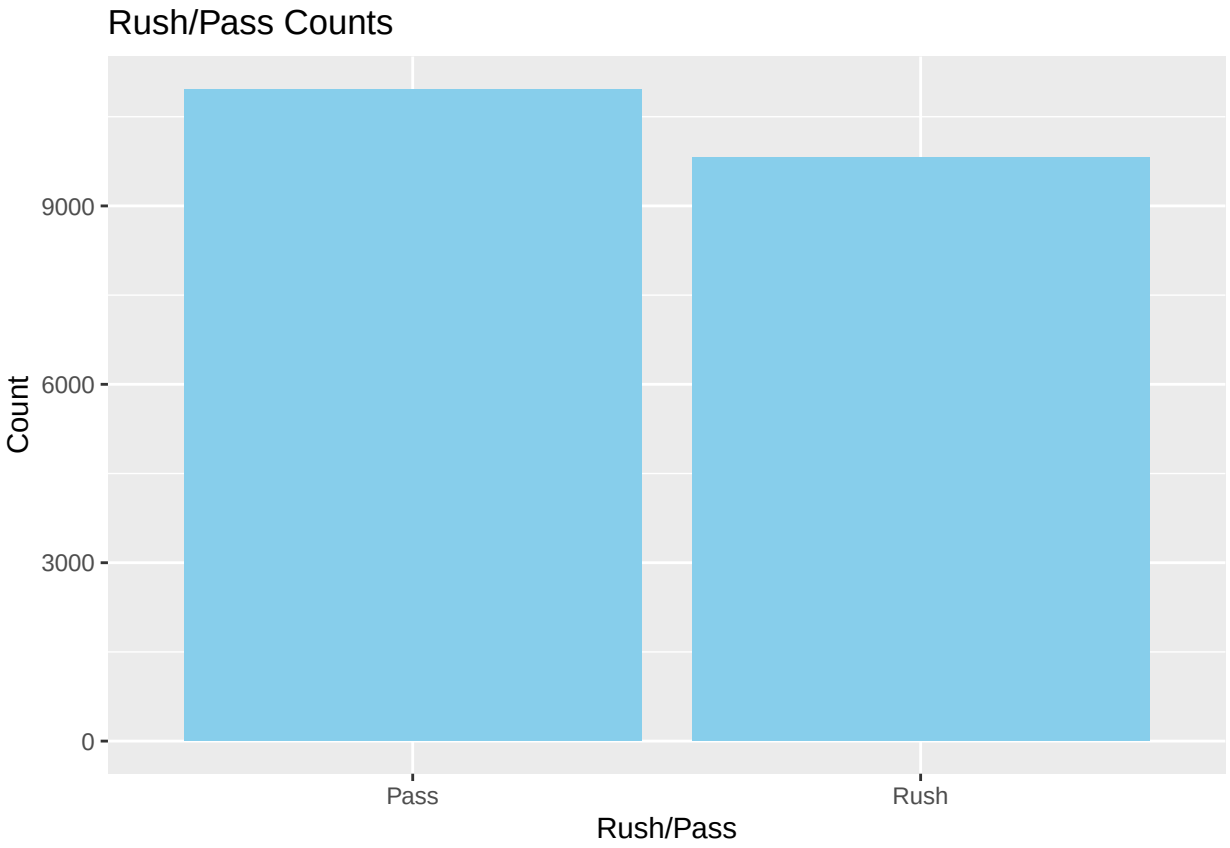

Figure 3: Rush Frequency Plot

Pass plays are ran more than rush plays in the red zone. This is suprising as it is generally assumed that rushing in the red zone is safer and an easier way to score.

Formation

Table 4: Formation Frequency

Formation	Count
SHOTGUN	12097
UNDER CENTER	6914
NO HUDDLE	437
NO HUDDLE SHOTGUN	1323
WILDCAT	2

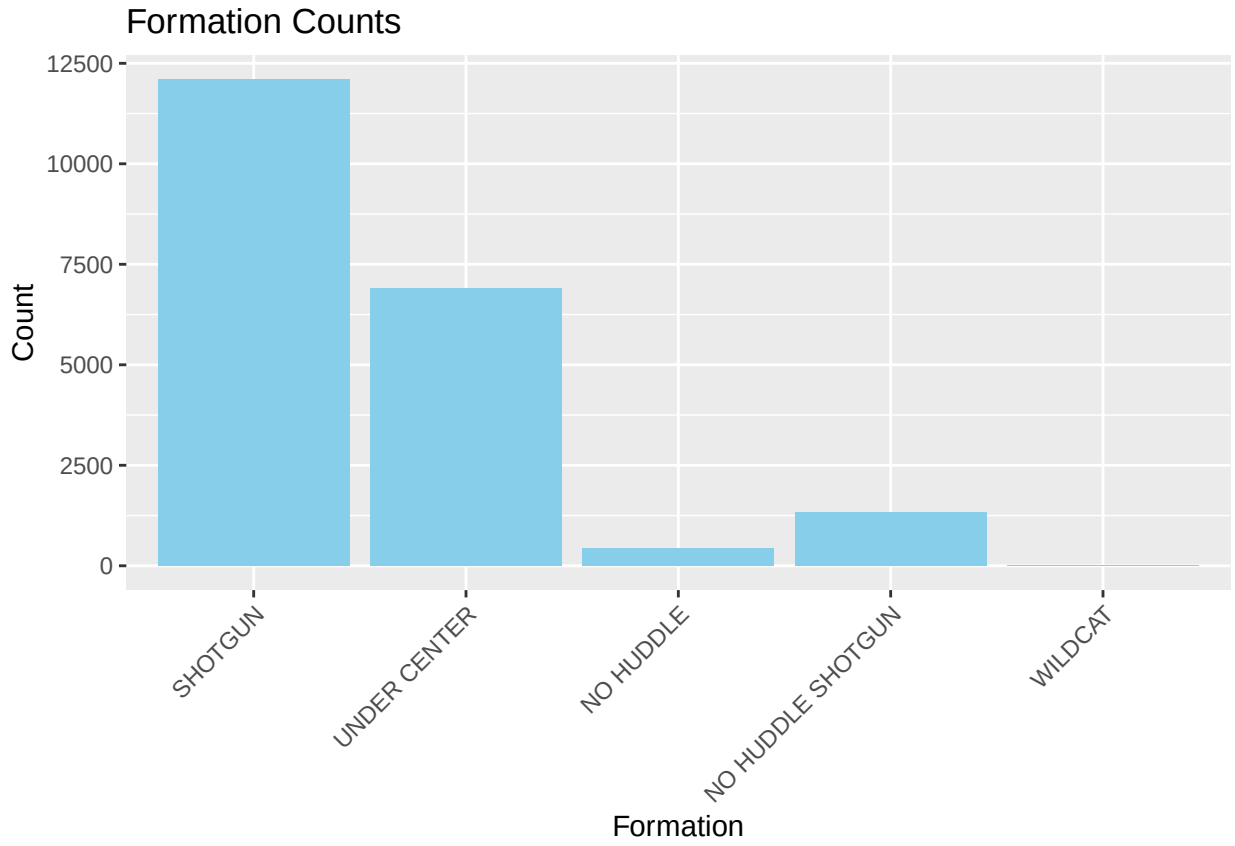

Figure 4: Formation Frequency Plot

There are two plays ran in the wildcat formation. It is not visible due to the larger counts of the other formations. Shotgun is the most common formation in the red zone, with under center being second. This is not suprising as these are the most common formations in football no matter what area of the field a team is positioned.

PlayType

Table 5: Play Type Frequency

Play Type	Count
PASS	10964
RUSH	9216
SCRAMBLE	593

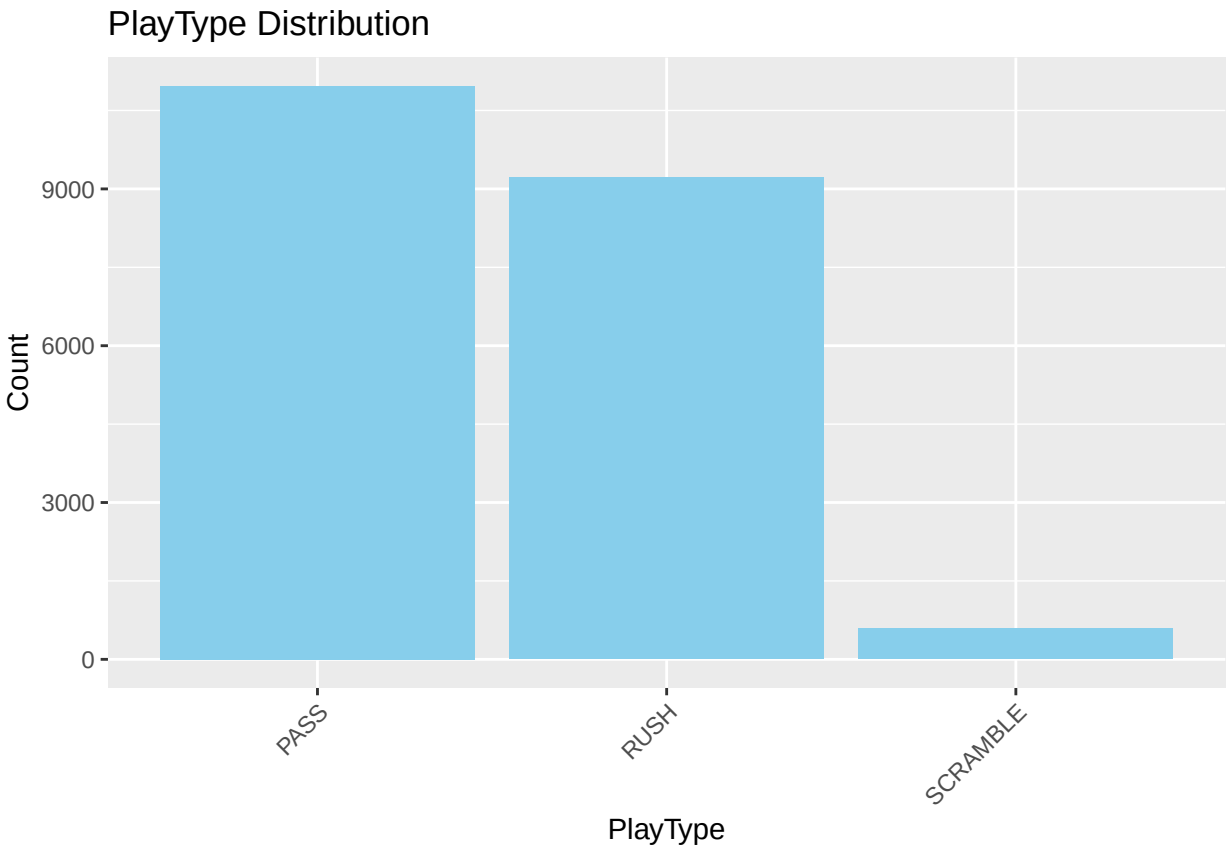

Figure 5: Figure 12: Play Type Frequency Plot

As observed earlier, more pass plays are ran in the red zone than rush plays. In this plot, it can be observed that scramble plays are much less common than the other two plays types. This is expected as scramble plays are generally a last resort by a quaterback to gain yards when a play has broken down.

Yards

```
Summary 1: Yards Gained/Lost Summary

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -20.000   0.000   2.000   3.327   5.000  19.000
```

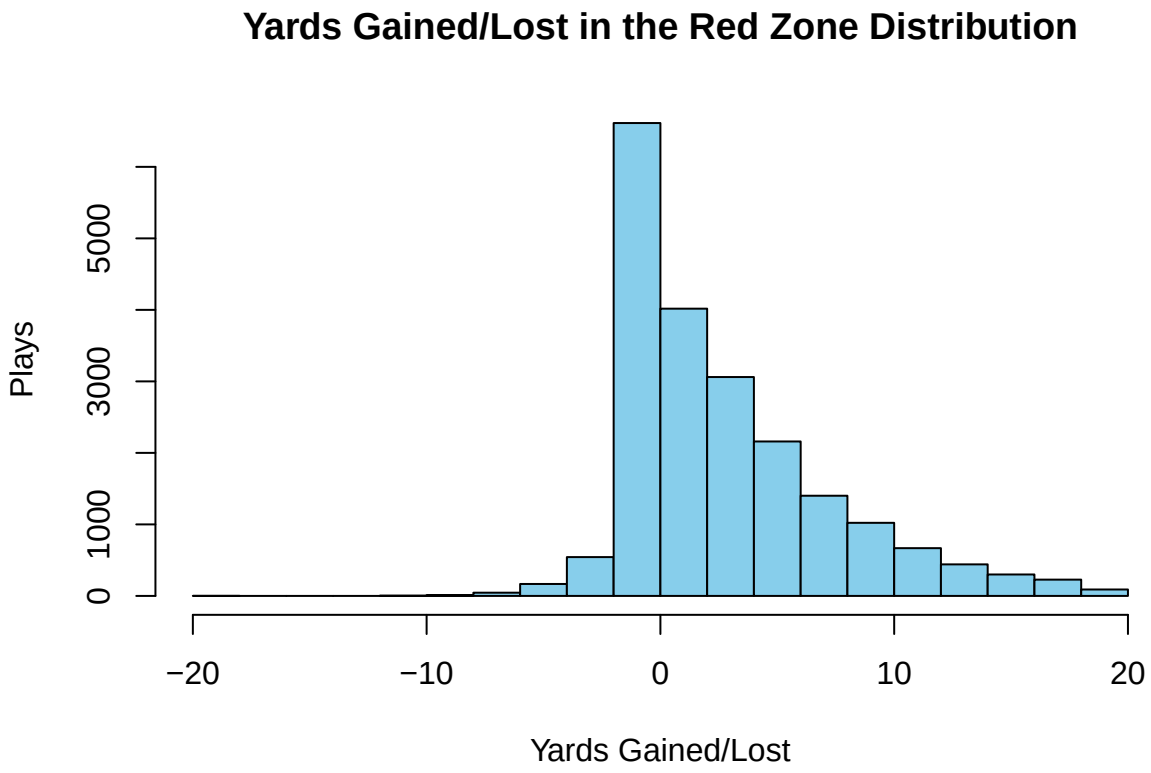

Figure 6: Yards Gained/Lost Distribution

It can be observed that the most common outcome for a play in the red zone is to lost around 1 or 2 yards, but uncommon to lost many more than that. In addition, when yards are gained, it is most commonly in increments of 1 to 5 yards.

3.0 Analysis

Four Techniques were used to carry out our analysis:

3.1 - Logistic Regression
3.2 - Random Forest Model
For each confusion matrix, note that the top labels are actual values (reference) and the labels on the left are predictions. ROC curves are used to compare the overall performance for the logistic regression models and the random forest models.
3.3 - Hypothesis Testing
3.4 - ANOVA and Post-Hoc Modeling

It was observed in figure/table 1 that there was mild imbalance in the IsTouchdown variable. Similarly, it was observed in figure/table 2 that there was also mild imbalance in the SeriesFirstDown variable. To account for this, undersampling was used to decrease the count of the majority class in each of these variables, respectively, to balance each of them for the logistic regression and random forest model training data. Since, there are already so many data points, we decided to use undersampling rather than oversampling. This improves the performance of our model without adjusting the minority class which is of interest. However, there are some potential drawbacks to undersampling, such as information loss and bias, which are accounted for in the conclusion section.

3.1 Logistic Regression Logistic regression is useful when dealing with multiple predictor variables and a binary outcome variable. By fitting logistic regression models, we can identify input variables that are significant in predicting if a play resulted in a touchdown or a first down, or not.

Related Guiding Questions - Are pass or run plays more likely to score a touchdown/achieve a first down) in the red zone? What formations are best for achieving these outputs in the red zone as well?

Formula for Input Variables = Formation + IsRush + IsPass

One logistic regression model will be used for the outcome variable IsTouchdown. The second logistic regression model will be used for the outcome variable SeriesFirstDown.

Cross validation is used to fit both models to reduce overfitting and produce the best possible model.

3.1.1 IsTouchdown Logisitic Regression Model: Significant Variables at 0.05 significance level:
- IsRush = "RUSH" - coefficient: -0.20196
Interpretation: If all other variables are constant, the log odds outcome variable of IsTouchdown will decrease by 0.20196 when IsRush = "RUSH".
- Formation = "UNDER CENTER" - coefficient: 0.13330
Interpretation: If all other variables are constant, the log odds outcome variable of IsTouchdown will increase by 0.13330 when Formation = "UNDER CENTER".

Table 6: Confusion Matrix for Logistic Regression Touchdown Prediction

	No Touchdown	Touchdown
No Touchdown	2049	489
Touchdown	2066	589

True Positives: 589/5193 = 11.3%
True Negatives: 2049/5193 = 39.5%
False Positives: 2066/5193 = 39.8%
False Negatives: 489/5193 = 9.4%
Sensitivity: 54.63% (Ability to predict minority level - poor)

Specificity: 49.79%
Overall Accuracy: 58% - POOR

3.1.2 SeriesFirstDown Logistic Regression Model:

Significant variables at 0.05 significance level:
- Formation = "UNDER CENTER" - coefficient: 0.13002
Interpretation: If all other variables are constant, the log odds outcome variable of IsTouchdown will increase by 0.13002 when Formation = "UNDER CENTER".
- Formation = "NO HUDDLE" - coefficient: 0.51728
Interpretation: If all other variables are constant, the log odds outcome variable of IsTouchdown will increase by 0.51728 when Formation = "NO HUDDLE".

Table 7: Confusion Matrix for Logistic Regression First Down Prediction

	No First Down	First Down
No First Down	2155	888
First Down	1475	675

True Positives: 675 / 5193 = 13%
True Negatives: 2155 / 5193 = 41.5%
False Positives: 1475 / 5193 = 28.4%
False Negatives: 888 / 5193 = 17.1%
Sensitvity: 43.19% (Ability to predict minority level - poor)
Specificity: 59.37%
Overall Accuracy: 55% - POOR

3.2 Random Forest

Random Forest Model - Random forest trees are useful when dealing with multiple predictor variables and a binary outcome variable. By creating random forest tree models, we can identify input variables that are significant in predicting if a play resulted in a touchdown or a first down, or not.

Guiding Questions - Are pass or run plays more likely to score a touchdown/achieve a first down) in the red zone? What formations are best for achieving these outputs in the red zone as well?

Formula for Input Variables = Formation + IsRush + IsPass

One random forest model will be used for the outcome variable IsTouchdown. The second random forest model will be used for the outcome variable SeriesFirstDown.

Cross validation is used to fit both models to reduce overfitting and produce the best possible model.

3.2.1 IsTouchdown Random Forest Model

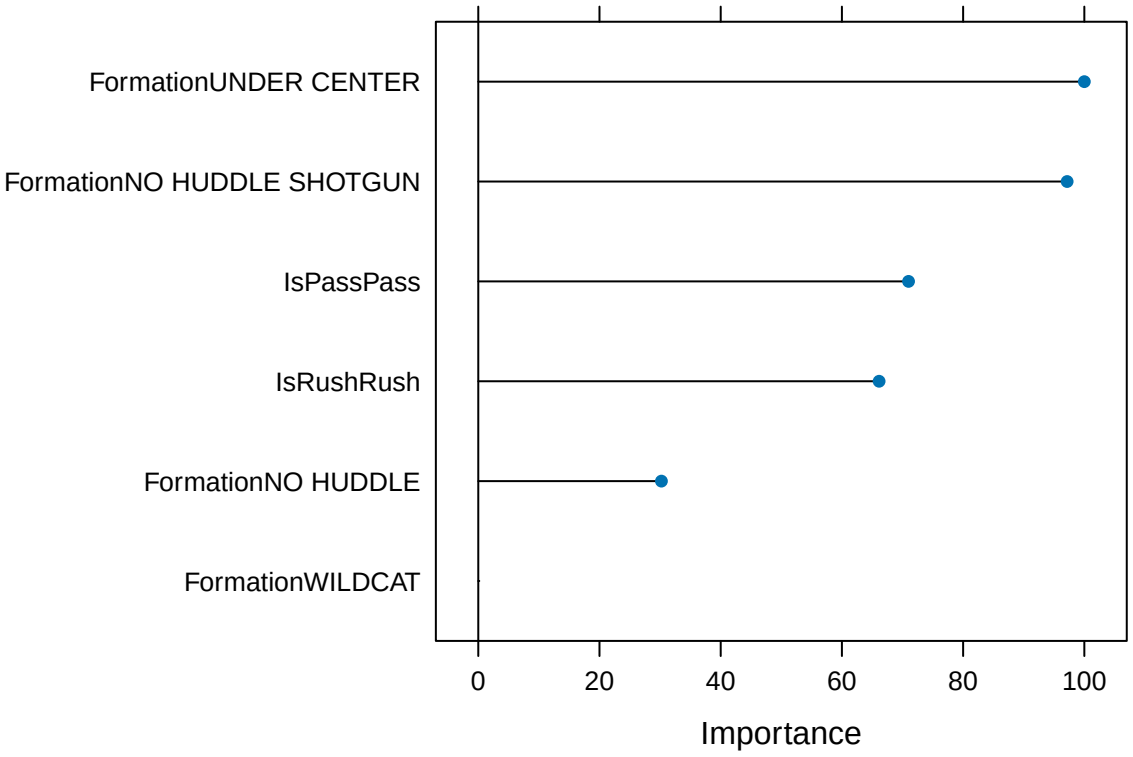

Figure 7: Random Forest Touchdown Predictor Importance

Formation = UNDER CENTER and Formation = NO HUDDLE SHOTGUN are found to be the most significant predictors of IsTouchdown. IsPass and IsRush seem to be somewhat significant but slightly less than the Formation variable.

Table 8: Confusion Matrix for Random Forest Touchdown Prediction

	No Touchdown	Touchdown
No Touchdown	1955	470
Touchdown	2160	608

True Positives: 608 / 5193 = 11.7%
True Negatives: 1955 / 5193 = 37.6%
False Positives: 2160 / 5193 = 41.6%
False Negatives: 470 / 5193 = 9.1%
Sensitvity: 56.4% (Ability to predict minority level - poor)
Specificity: 47.5%
Overall Accuracy: 50%

3.2.2 SeriesFirstDown Random Forest Model

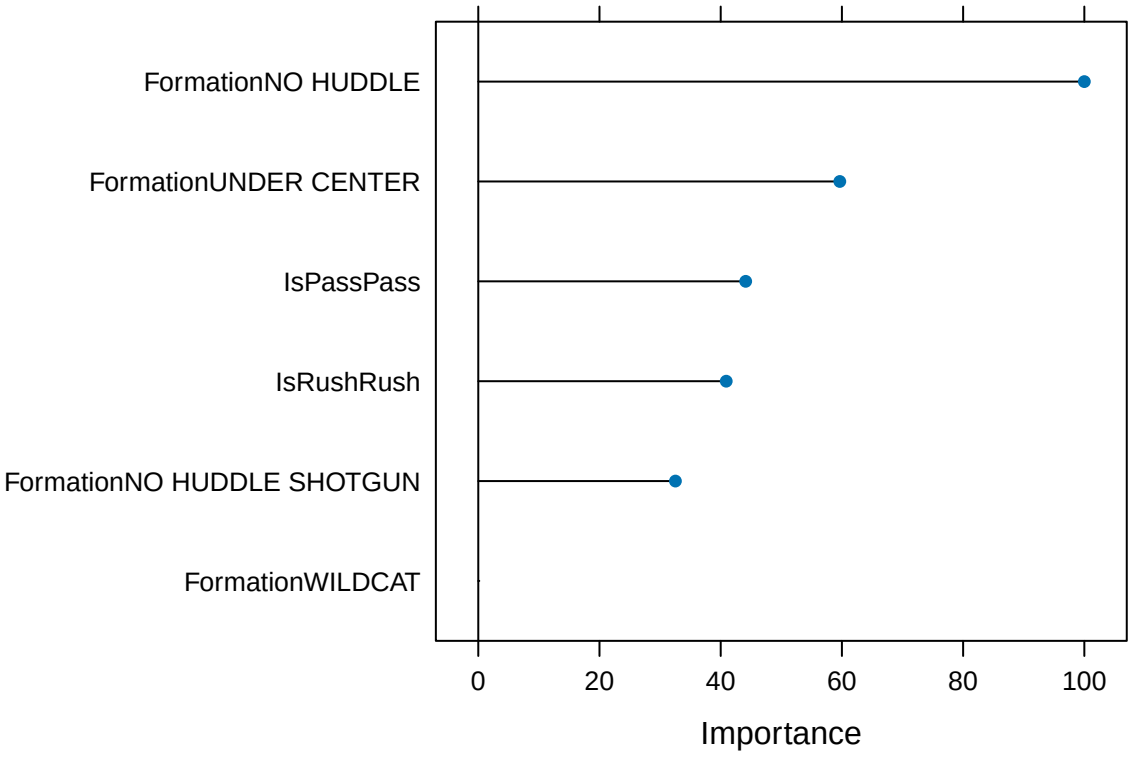

Figure 8: Random Forest First Down Predictor Importance

Formation = NO HUDDLE is found to be the most significant predictor of SeriesFirstDown. Formation = UNDER CENTER, IsPass, and IsRush are less significant, though still of somewhat importance.

Table 9: Confusion Matrix for Random Forest First Down Prediction

	No First Down	First Down
No First Down	2292	948
First Down	1338	615

True Positives: 615 / 5193 = 11.8%
True Negatives: 2292 / 5193 = 44.1%
False Positives: 1338 / 5193 = 25.8%
False Negatives: 948 / 5193 = 18.3%
Sensitvity: 39.3% (Ability to predict minority level - poor)
Specificity: 63.1%
Overall Accuracy: 56% - POOR

ROC Curve IsTouchdown

```
## Area under the curve (AUC): 0.522
```

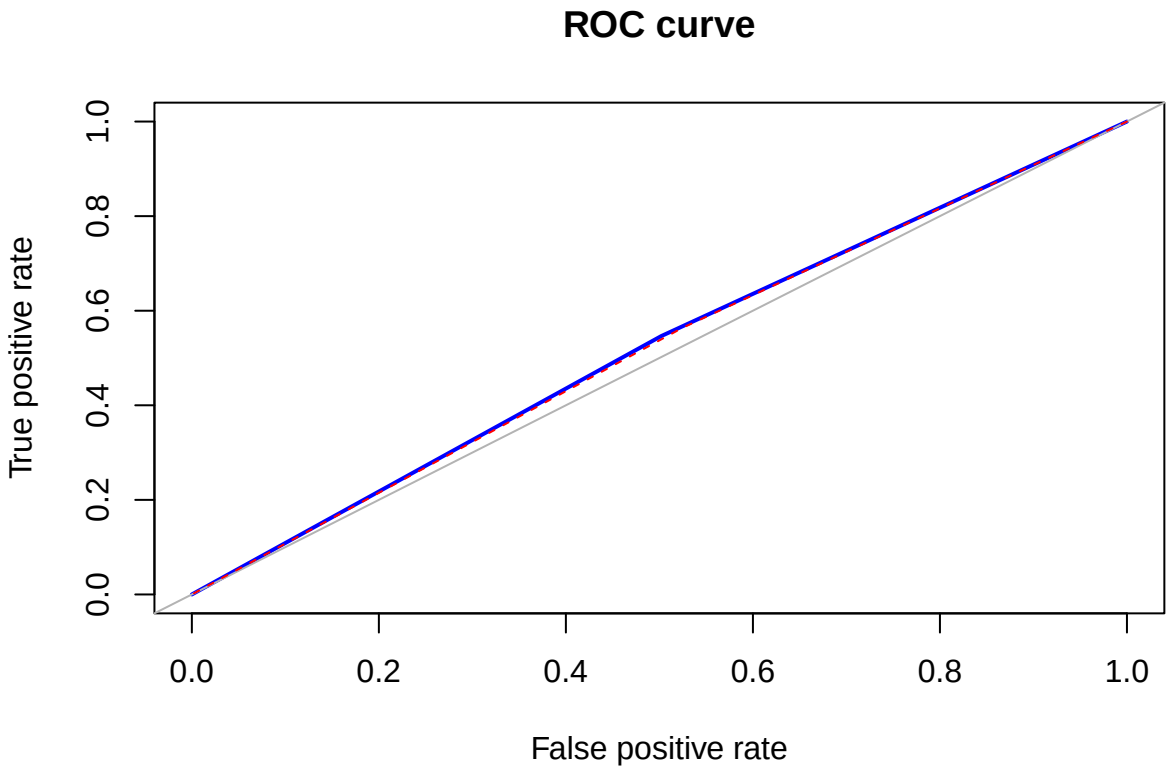

Figure 9: ROC Curve for Logistic Regression and Random Forst Touchdown Predicition Performance

```
## Area under the curve (AUC): 0.520
```

(Blue) Logistic Regression AUC = 0.522
(Red) Random Forest AUC = 0.520
The two models essentially have the same performance (both poor).

ROC Curve SeriesFirstDown

```
## Area under the curve (AUC): 0.513
```

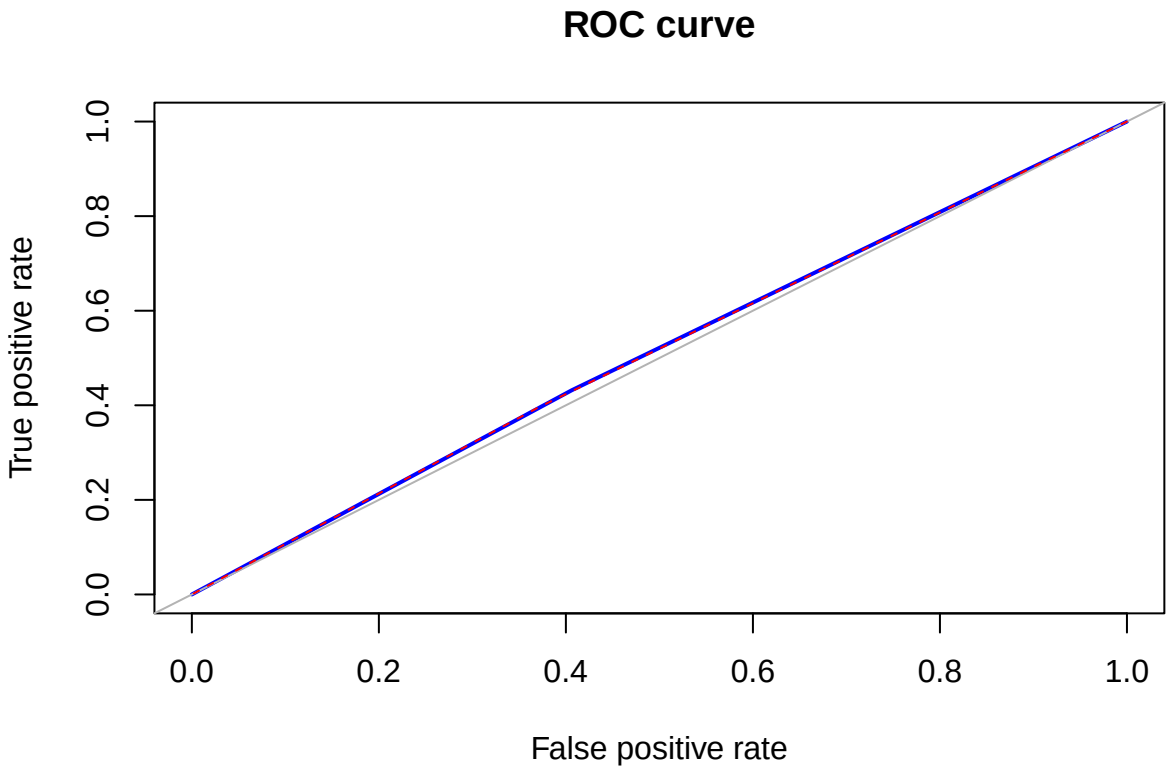

Figure 10: ROC Curve for Logistic Regression and Random Forst First Down Predicition Performance

```
## Area under the curve (AUC): 0.512
```

(Blue) Logistic Regression AUC = 0.513
(Red) Random Forest AUC = 0.512 The two models essentially have the same performance (both poor).

3.3 Hypothesis Testing for Yards Gained/Lost

Hypothesis testing - useful for determining if their is a statisically significant difference in the means of two samples

Guiding Questions - Are pass or run plays more likely to gain yards in the red zone? What formations are best for gaining yards?

First two-sample hypothesis test will determine if their is a difference between the mean yards gained on pass plays vs rush plays.

Input Variables: PlayType (rush = "RUN" + "SCRAMBLE", pass = "PASS") Output Variable: Yards

The second two-sample hypothesis test will determine if their is a difference between the mean yards gained on plays from a UNDER CENTER formation vs SHOTGUN plays, as these are the most common formation types in the red zone.

3.3.1 Rush vs. Pass Two-Sample Hypothesis Test

Rush vs Pass:

P-Value: 2.2e-16

Interpretation: The true difference in mean yards gained between pass and rush plays is not equal to 0 (rejection of null hypothesis).

Mean of rush yards: 2.942196
Mean of pass yards: 3.672109

Passing plays gain more yards than rushing plays in the red zone.

3.3.2 Shotgun vs. Under Center Two-Sample Hypothesis Test

Formation:

P-Value: 2.2e-16

Interpretation: The true difference in mean yards gained between plays ran in an under center formation and plays run in a shotgun formation is not equal to 0 (rejection of null hypothesis).

Under center mean yards gained = 2.773358
Shotgun mean yards gained = 3.609655

Plays ran from a shotgun formation gain more yards than plays ran from an under center formation in the red zone.

3.4 ANOVA and Post-Hoc Testing for Yards Gained/Lost

ANOVA and Post-Hoc - useful when determining if the means of two or more groups are different from each other in a statistically significant manner. This is another way of comparing if pass plays and run plays are different from each other in terms of yards gained in the red zone, as well as how formations compare in terms of yards gained.

Guiding Questions - Are pass or run plays more likely to gain yards in the red zone? What formations are best for gaining yards?

First Anova and Post-Hoc Test - Input = Formation, Output = Yards

Second Anova and Post-Hoc Test - Input = PlayType, Output = Yards

3.4.1 Formation ANOVA and Post-Hoc

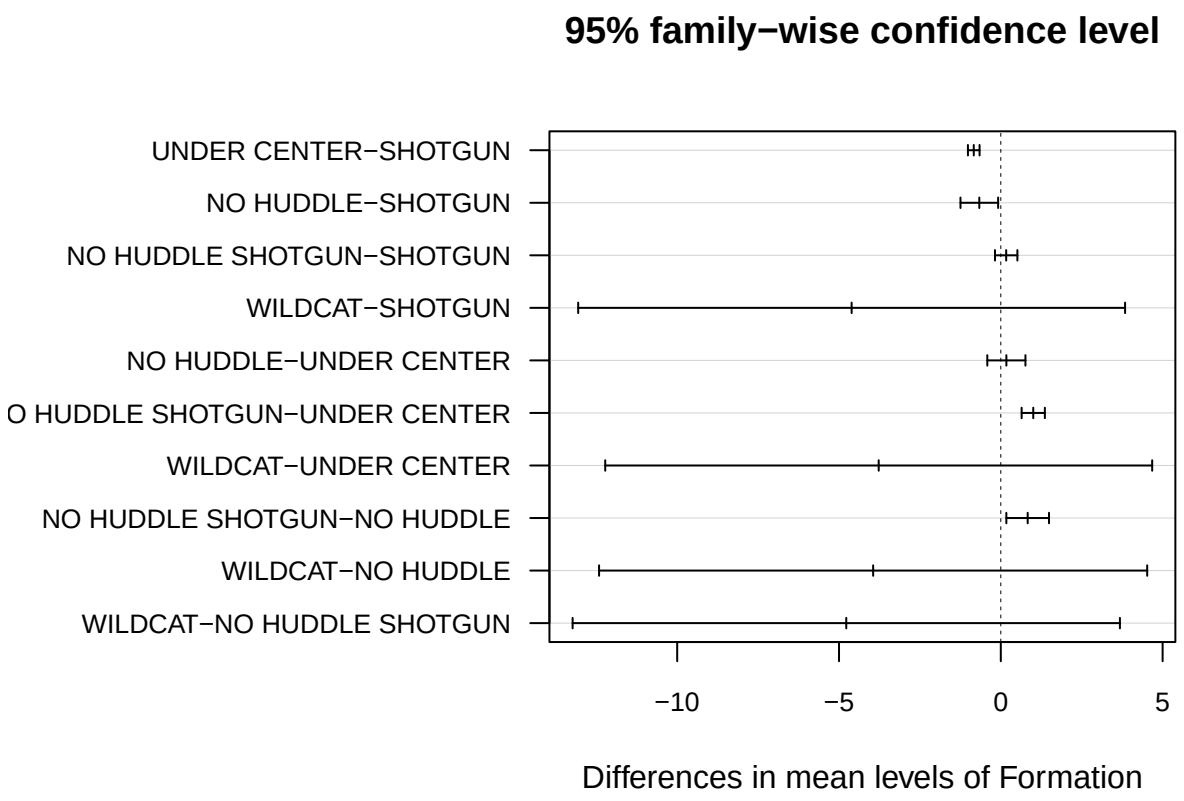

Figure 11: ANOVA Model for Yards Gained/Lost by Formation

The difference in mean yards gained for some formations in the red zone were not statistically significant. SHOTGUN was shown to result in a greater mean yards gained than UNDER CENTER and NO HUDDLE. In addition, plays ran from UNDER CENTER were shown to gain less yards than plays ran from NO HUDDLE SHOTGUN (NO HUDDLE SHOTGUN = the offensive team did not huddle up to discuss the play beforehand and lined up in a shotgun formation).

3.4.2 Play Type ANOVA and Post-Hoc

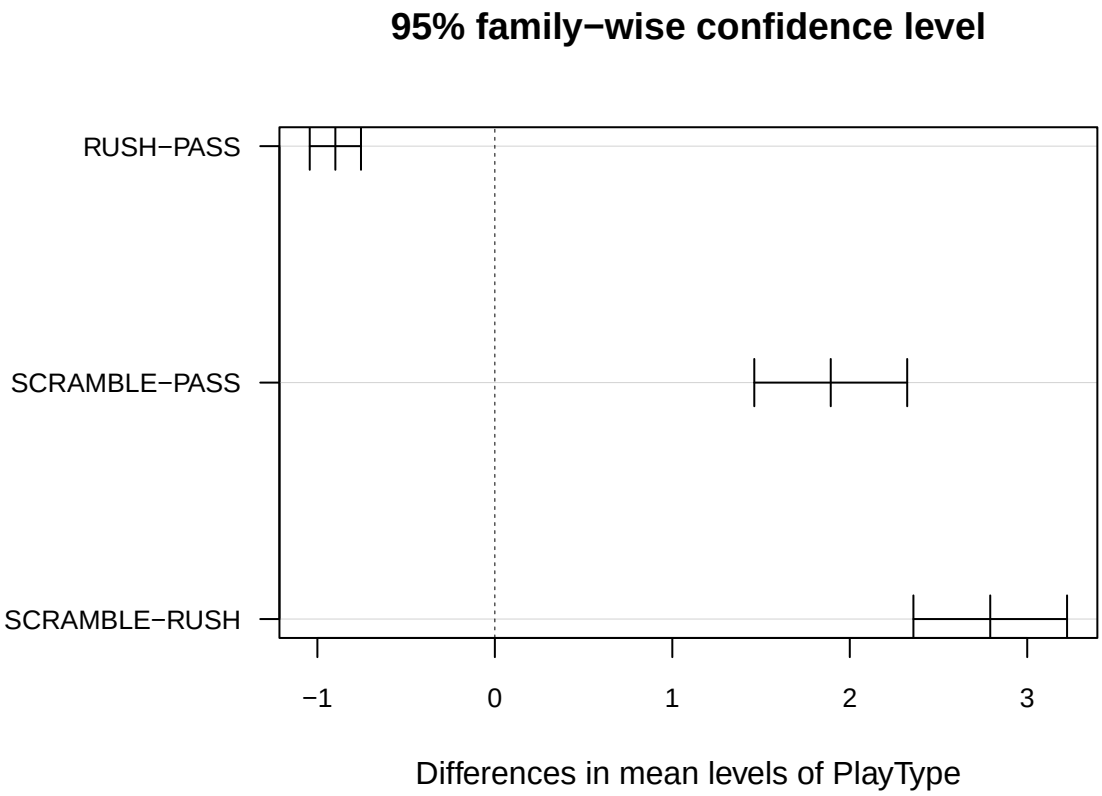

Figure 12: ANOVA Model for Yards Gained/Lost by Play Type

It can be observed that the mean yards gained between rush plays, pass plays, and scramble plays, are each significantly different in the red zone. Pass plays are shown to gain more yards than rush and scramble plays. In addition, scramble plays are shown to gain more yards than rush plays.

4.0 Conclusions

4.1: IsTouchdown and SeriesFirstDown

Both the logistic regression and random forest models were ineffective in predicting each of the outcome variables IsTouchdown and SeriesFirstDown, respectively. The logistic regression models had a prediction accuracy of 58% for a touchdown and 55% for a first down. The random forest models had a prediction accuracy of 50% for a touchdown and 56% for a first down. This could be due to the fact that the predictors used in the models were not significantly correlated in a suffienct way to predict these outcome variables.

The logistic regression model and the random forest model found Formation = UNDER CENTER to be significantly correlated with the IsTouchdown variable, and the logistic regression model showed a positive correlation. The random forest model also found Formation = NO HUDDLE SHOTGUN to be significant. In contrast, a play being a pass or a rush was significant in the logistic regression model, but slightly less significant in the random forest model. Furthermore, a play being a rush was shown in the logistic regression model to have a negative effect on the probability a play was a touchdown. However, these results should be taken with caution do to the inaccurate prediction model.

In terms of the SeriesFirstDown variable, Formation = UNDER CENTER and Formation = NO HUDDLE were both found to be the most significant variables. In addition, from the logistic regression model, Formation = NO HUDDLE was positively correlated with a play resulting in a first down, while Formation = UNDER CENTER was as well, though less so. However, these results should be taken with caution do to the inaccurate prediction model.

4.2: Yards

From the two-sample hypothesis test, it was shown that the true difference in mean yards gained between plays ran in an under center formation and plays run in a shotgun formation is not equal to 0. It was found that plays ran from a shotgun formation resulted in a mean 3.6 yards gained, whereas plays from under center resulted in a mean 2.8 yards gained. This is reinforced by the ANOVA model. However, the ANOVA and post-hoc model included more variations of formations as well. The difference in mean yards gained for some formations were not statistically significant. SHOTGUN was shown to result in a greater mean yards gained than UNDER CENTER and NO HUDDLE. In addition, plays ran from UNDER CENTER were shown to gain less yards than plays ran from NO HUDDLE SHOTGUN.

In terms of whether a play was a pass or rush, this difference was also shown to be statistically significant from the two-sample t-test. It was found that rush plays resulted in a mean 2.9 yards gained, whereas pass plays resulted in a 3.67 yards gained. This is in agreement with the ANOVA and post-hoc test, through which it can be observed that the mean yards gained between rush plays, pass plays, and scramble plays, are each significantly different. Pass plays are shown to gain more yards than rush and scramble plays. In addition, scramble plays are shown to gain more yards than rush plays.

4.3 Optimizing Red Zone Offense

Our findings show that in an attempt to score a touchdown in the red zone, an offensive team in the NFL should line up in an under center formation. The play type was not found to be significant. In an attempt to achieve a first down in the red zone, an offensive team should not huddle up before a play, or line up in an under center formation. The play-type was once again not found to be significant. Lastly, to have the best chance to gain yards in the red zone, an offensive team should line up in a shotgun formation, huddle up to discuss the play or not, and pass from this formation.

Another way to display this:

To score a touchdown - Formation = UNDER CENTER
To get a first down - Formation = NO HUDDLE or Formation = UNDER CENTER
To gain yards - Formation = SHOTGUN or NO HUDDLE SHOTGUN, PlayType = Pass

(The variables positively associated with scoring a touchdown or getting a first down should be taken with caution, as shown by the inaccurate prediction models.)

4.4: Limitations and Biases

- Data had to be manipulated due to imbalance. Data was balanced using undersampling for analysis and therefore information was lost and bias may have been introduced.
- Logistic regression and random forest models may have been biased toward the majority class.
- This idea is reinforced by inaccurate prediction.

5.0 References

The data was downloaded as csv files from this source: https://nflsavant.com/about.php